Le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication

Actes de conférence

6ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ
ⴷ ⵜⵉⴽⵏⵓⵍⵓⵊⵉⵢⵜ ⵏ ⵓⵏⵖⵎⵉⵙ ⴷ ⵓⵎⵢⴰⵡⴰⴹ

الأمازيغية
وتكنولوجيا المعلومات والتواصل

Technologies d’Information et de Communication pour l'Amazighe

Coordination

ATAA ALLAH Fadoua

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵉⴽⵏⵓⵍⵓⵊⵉⵢⵜ ⵏ ⵓⵏⵖⵎⵉⵙ ⴷ ⵓⵎⵢⴰⵡⴰⴹ

الأمازيغية وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication
pour l'Amazighe

Centre des Etudes Informatiques,
des Systèmes d'Information et de Communication

Actes de conférence

6ème Conférence Internationale

ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵉⴽⵏⵓⵍⵓⵊⵉⵢⵜ ⵏ ⵓⵏⵖⵎⵉⵙ ⴷ ⵓⵎⵢⴰⵡⴰⴹ

الأمازيغية وتكنولوجيا المعلومات والتواصل

Technologies d'Information et de Communication pour l'Amazighe

Coordination
Ataa Allah Fadoua

Publications de l'Institut Royal de la Culture Amazighe
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Série : Colloques et séminaires N° 42

Titre
Technologies d'Information et de Communication pour l'Amazighe

Coordination
Ataa Allah Fadoua

Conception
Nadia Kiddi (Unité de l'édition)

Editeur
Institut Royal de la Culture Amazighe

Imprimerie
El Maârif Al Jadida - Rabat

Dépôt légal
2015 MO 1163

ISBN
978-9954-28-184-0

ISSN
2421-9711

PREFACE

Les technologies d'information et de communication constituent un enjeu considérable pour la consolidation des politiques linguistiques, visant la préservation et la promotion des langues et cultures peu dotées informatiquement. D'où l'importance de mettre en avant le gain que peuvent apporter les nouvelles technologies à ces langues et cultures, en focalisant sur les démarches à entretenir et les travaux de recherche à mener à cette fin. Dans ce contexte que le Centre des Etudes Informatiques, des Systèmes d'Information et de Communication (CEISIC) de l'Institut Royal de la Culture AMazighe (IRCAM) organise bi-annuellement, depuis 2004, la conférence internationale sur les Technologies d'Information et de Communication pour l'AMazighe (TICAM).

Le présent ouvrage présente les actes de la 6ème édition de cette conférence tenue à Rabat les 24 et 25 novembre 2014, au siège de l'IRCAM. Il rassemble en plus du mot d'accueil prononcé par M. BOUKOUSS Ahmed, le Recteur de l'IRCAM, une vingtaine d'articles qui ont été sélectionnés suite à une évaluation assurée par au moins deux spécialistes. Ces articles s'articulent autour d'un ensemble de thèmes dont nous citons : l'enseignement des langues assisté par ordinateur, les ressources et outils linguistiques, le traitement de la parole et la reconnaissance optique des caractères.

En outre des communications publiées au sein de cet ouvrage, la conférence a permis aux participants venus aussi nombreux d'écouter trois conférenciers invités, de renommée nationale et internationale. Ils ont partagé avec eux leur savoir et expertise, autour des thèmes d'actualité et en rapport avec l'objectif principal de la conférence. Ces conférenciers sont M. CHOUKRI Khalid, Directeur Général de l'agence ELDA, M. PARADOWSKI Michal, enseignent chercheur à la faculté de la linguistique appliquée de l'université de Varsovie, et M. LEFDAOUI Youssef, enseignent chercheur à l'Ecole Supérieure de Technologie de Salé.

Et c'est à travers ces quelques lignes que nous souhaitons remercier chaleureusement nos invités pour l'honneur qu'ils nous ont fait par leur participation et pour les exposés qu'ils nous ont présentés. Nous tenons à remercier également les membres du Comité de Lecture pour la qualité et la rigueur de leur travail et le sérieux qu'ils ont apporté en vue d'améliorer la qualité des soumissions. Nous adressons aussi nos sincères remerciements au Comité d'organisation et à tous ceux qui ont œuvré et contribué pour la réussite de cette édition.

Comité scientifique

- Aboutajdine Driss (FSR, Rabat)
- Ait Ouguengay Youssef (IRCAM, Rabat)
- Aoughlis Farida (UMMTO, Tizi Ouzou)
- Ansar Khalid (IRCAM, Rabat)
- Ataa Allah Fadoua (IRCAM, Rabat)
- Bennani Samir (EMI, Rabat)
- Boulaknadel Siham (IRCAM, Rabat)
- Bouyakhf Houssaine (FSR, Rabat)
- Cavalli Sforza Violetta (AUI, Ifrane)
- Cherkaoui Chihab Dine (ENCG, Agadir)
- El Hamdani Abdelfettah (IERA, Rabat)
- El Mouradi Abdelhak (ENSIAS, Rabat)
- El qadi Abderrahim (EST, Meknes)
- Fadouli Nouredine (EMI, Rabat)
- Fakir Mohamed (FST, Beni Mellal)
- Idrissi Najlae (FST, Beni Mellal)
- Khalidi Idrissi Mohamed (EMI, Rabat)
- Maach Abdelilah (EMI, Rabat)
- Mammas Driss (EST, Agadir)
- Modi Issouf (MEN, Niger)
- Pognan Patrice (INALCO, Paris)
- Rachidi Ali (ENCG, Agadir)
- Rami Salim (FPS, Safi)
- Roche Mathieu (Univ-Montpellier 2, Montpellier)
- Rosso Paolo (UPV, Valence)
- Silberztein Max (Univ-Franche-Comté, Besançon)
- Souifi Hamid (IRCAM, Rabat)
- Tigziri Noura (UMMTO, Tizi Ouzou)
- Yousfi Abdellah (FSJES, Rabat)
- Zenkouar Lahbib (EMI, Rabat)
- Zock Michael (Univ-Marseille, Marseille)

Comité d'organisation

- Ait ouguengay Youssef
- Ataa Allah Fadoua
- Boulaknadel Siham
- Boumediane Mounia
- EL Hamdaoui Amal
- El Marssi Karim
- Frain Jamal

Mot du Recteur

Chers collègues,

Il m'est particulièrement agréable de vous accueillir au nom de l'IRCAM à l'occasion de la 6e édition du TICAM consacrée à l'apport des nouvelles technologies à la mise à niveau de l'amazighe et à sa promotion dans le contexte de sa constitutionnalisation et de son institutionnalisation en tant que langue officielle aux côtés de la langue arabe. Ce nouveau statut de l'amazighe est porteur d'enjeux inédits, et dont le principal est sans doute la capacitation de l'amazighe pour qu'il joue pleinement les fonctions qui lui sont dévolues par la Constitution, notamment dans les domaines de l'éducation et de la communication dans la sphère publique. Dans cette perspective, le manque de vision stratégique allié aux moyens traditionnels de l'enseignement-apprentissage, notamment l'enseignement en présentiel, le centrage sur l'enseignant, l'utilisation du tableau noir et des supports pédagogiques en papier, l'usage d'approches, de méthodes et de démarches changeantes, etc. se sont révélés coûteux et parfois inefficaces. Ces moyens caducs ont conduit à des résultats peu probants pour ne pas dire qu'ils ont contribué à la « faillite » de l'enseignement jugée à l'aune des évaluations nationales comme celle du PNEA, et des évaluations internationales comme celles de PIRLS et de TIMMS.

L'enseignement-apprentissage de l'amazighe a tout à gagner en tirant les leçons qui s'imposent de cette situation, et ce en se modernisant grâce à l'usage adéquat des nouvelles technologies de l'information et de la communication appliquées à l'éducation. Les NTICE offrent aujourd'hui une panoplie de produits software comme les plateformes numériques de formation en ligne ouverte à tous (MOOCs), les équipements mobiles, les tableaux interactifs, les applications de tous ordres qui révolutionnent les sciences de l'éducation et les pratiques pédagogiques et didactiques. Ces technologies, contrairement à ce que d'aucuns pourraient penser, peuvent contribuer à démocratiser l'éducation, et à rendre plus performant notre système d'éducation-formation. En tout cas, l'enseignement de l'amazighe et par l'amazighe, aussi bien dans le formel que dans le non formel et dans la lutte contre l'analphabétisme, ne peut que bénéficier des apports des NTICE. A cet égard, je voudrais rendre hommage aux efforts consentis par les collègues du CEISIC et des autres centres de recherche, et aux chercheurs externes à l'IRCAM, qui ont produit des CD éducatifs pour l'apprentissage de la langue amazighe, de la graphie tifinaghe et des éléments de la culture. L'innovation pédagogique passe donc par l'innovation technologique. C'est une évidence. D'où la pertinence et l'importance de cette rencontre.

Merci à vous, toutes et tous, de vous impliquer dans ce vaste chantier en acceptant l'invitation de l'IRCAM à cette édition du TICAM.

Table des matières

TICAM'14, p. 11-23

Towards an Adaptive Platform for Supporting Language Learning through Reading

Violetta Cavalli-Sforza **Idriss Agoumi**
Al Akhawayn University
v.cavallisforza@aui.ma idriss.agoumi@gmail.com

Abstract

We present ARA, a platform for supporting reading as one of the fundamental activities in language learning. ARA consists of integrated tools, connected through a database rich in linguistic information, that can be configured and reused to satisfy the needs of different teachers and students and will, in future work, support adaptive delivery of reading materials.

1. Motivation and Related Work

Reading is one of the four fundamental skills targeted by foreign language instruction. Reading and related activities are essential for learning vocabulary and its proper use in context, idiomatic expressions, grammar and virtually all other aspects of written language. The power of reading in one's native language, as well as in foreign language learning, has been extensively advocated by several researchers, and in particular Stephen Krashen (20004-a; 2004-b; 2013). Texts that provide *comprehensible input* to the learner —language input that is just beyond where the learner currently is—provide an opportunity for incidental learning of vocabulary, in addition to practicing explicitly targeted vocabulary (intentional learning).While there is some discussion as to which type of learning is most effective for language learning, both are part of a well-balanced language program (Waring & Nation, 2004).

Tools for supporting reading are particularly important for languages where the written form deviates from the commonly spoken forms and where complex morphology makes words used in context appear sometimes quite different from their dictionary lookup form. Arabic, with its spoken dialectal varieties and its formal modern written "standard" form is an example of such a language, though it is by no means unique.

We presents the Arabic Reading Assistant (ARA) platform—so called because, though not limited to Arabic in any way, the project was initially motivated by Arabic—a platform aimed at supporting reading and designed as a collection of integrated tools, connected through a database rich in linguistic information, that can be adapted and reused to satisfy the needs of different teachers and students and will, in future work, support adaptive delivery of reading materials.

The last ten years have witnessed the development of a large number of tools and platforms in support of computer-assisted learning in both commercial and academic/research spheres. Some of these tools, such as the Learning Management Systems (LMS) are very general; others are targeted at teaching very specific subjects, spanning the range from the humanities to science and math. Many systems have also been developed for language learning, some combining sophisticated multimedia capabilities and/or speech technology, others focusing more on providing intelligent response to users. Among the large variety of present offerings, three have most closely influenced our thinking and design and are briefly reviewed below.

The first and oldest of the three is GLOSSER-RUG, a computer-assisted language learning (CALL) system, designed to help Dutch native people (with some background in French) learn more about French Morphology. Our interest in studying this system was mainly for its morphological analysis approach. The main modules of GLOSSER-RuG are: a Part-Of-Speech (POS) disambiguator, a morphological analyzer, a dictionary lookup and a corpora search. The disambiguator is applied before morphological analysis in order to obtain the most plausible morphological analysis in context and get the right word from the dictionary lookup. The user, while reading a text, can select an unknown or unfamiliar word; the program makes available the internal structure of the word, including the grammatical information encoded in morphology, the dictionary entry of the word and examples of the word from corpora (Nerbonne & Smit, 1996)[1].

A second project that is closely related to ARA is ARET (Arabic Reading Enhancement Tools), funded by the U.S. Department of Education and based at the Linguistic Data Consortium (LDC) (Maamouri *et al*., 2012). Specifically targeted at Arabic, it includes two tools: 1) the Arabic Reading Facilitation Tool (ARFT), and 2) the Arabic Reading Assessment Tool (ARAT). ARFT provides the user with direct access to a set of readings, taken from the Al-Kitaab textbook (Brustad *et al*., 2004) and organized by volume, chapter and page number. The user can click on any word in the passage displayed to bring up the full morphological analysis and gloss for the word in context, along with a glossary entry for the associated lemma, and a summary of other Arabic citation forms that are related by root. It adds a text-to-speech system that gives the pronunciation of individual words and extended text. ARAT supports the creation of tests and quizzes for specific Arabic reading skills. It provides registration of faculty and student user accounts, creation of courses with schedule plans and content-based resources, creation, presentation and scoring of tests and quizzes; and overall record-keeping of resources, activities and test scores.

The third system, which addresses specifically student-adapted presentation of learning materials, is project REAP (Reader-Specific Lexical Practice for Improved Reading Comprehension), also funded by the U.S. Department of Education and based at Carnegie Mellon University (Brown & Eskenazi, 2004). The original concept of this project resides in building a student model based on learned vocabulary. Each student must learn a certain number of words appearing in a list of target words. REAP retrieves text materials adapted to the student's level, each of which containing short passages in a specific context[2]. The core ideas of the REAP project are: 1) a search engine that finds text passages satisfying

1 GLOSSER-RuG is accessible via http://www.let.rug.nl/glosser/Glosser/.
2 REAP is accessible via http://reap.cs.cmu.edu/#Demos.

very specific lexical constraints; 2) selecting materials from an open-corpus, thus satisfying a wide range of student interests and classroom needs; and 3) the ability to model an individual degree of acquisition and fluency for each word in a constantly-expanding lexicon so as to provide student-specific practice and remediation. REAP was initially developed for learners of English as a second/foreign language, and later extended to French and Portuguese. It currently also incorporates not just vocabulary but also some aspects of grammar.

2. Structure and Contents of ARA

In this section, we describe the general structure of the ARA platform and provide some details about specific components. Later sections focus on the distinguishing aspects of the platform. ARA currently exists as a desktop system, a collection of interface developed in C# and interacting through a MySQL database.

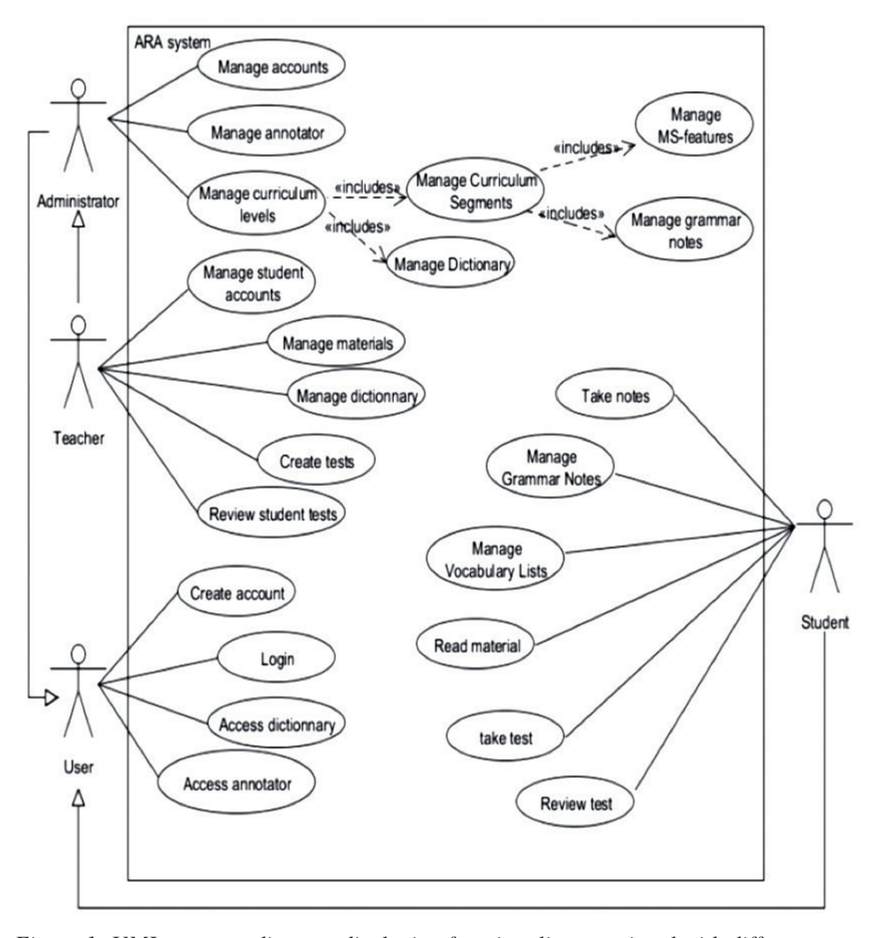


Figure 1: UML use case diagram displaying functionality associated with different users of the platform: administrator(s), teachers and students

Access to the platform's functionality is by role: administrator, teacher or student. The three roles require separate logins but the interfaces share the same look and feel provided by the C# Dock Container library. Figure 1 provides an overall view of the use cases for the system. We describe briefly the functionality associated with the three roles.

2.1. Administrator Functionality

The administrator plays the role of a "super teacher" or course coordinator, controlling the organization of instructional materials and instructors. In addition to traditional administrative functions (e.g. account management), one of the administrator's main roles is structuring the curriculum, which includes setting up the different curriculum levels, the dictionary entries and segments associated with each level and the target grammatical concepts associated with each segment. The latter are represented in Figure 1 by the entries Manage MS-features and Manage Grammar Notes.

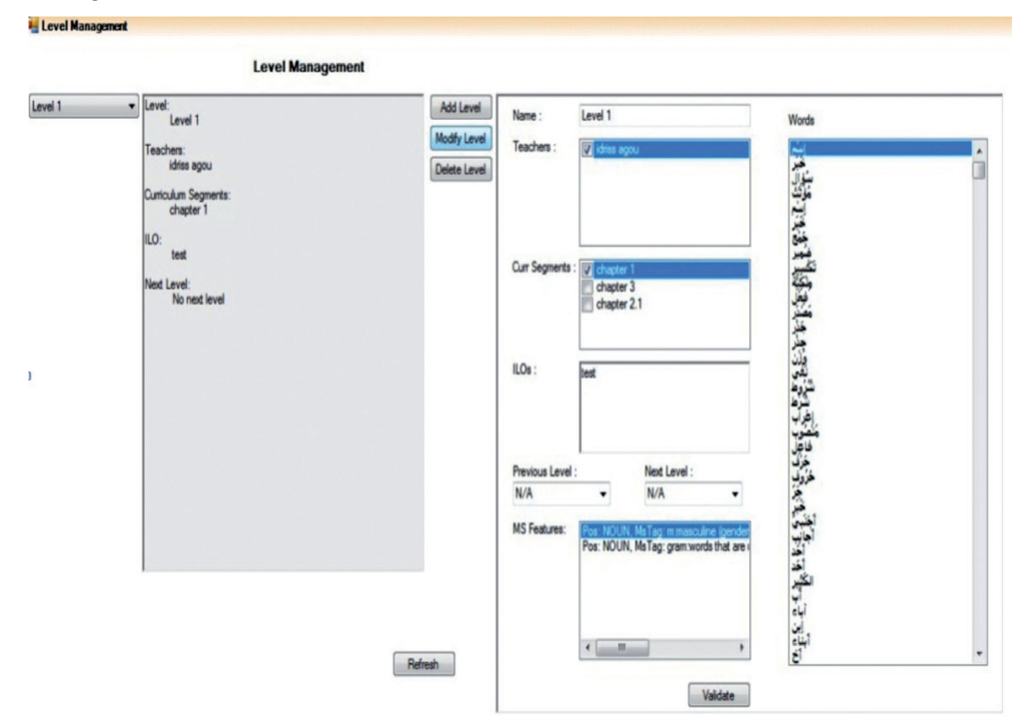


Figure 2 : Administrator interface

Figure 2 displays one of the interfaces through which the administrator acts to structure the curriculum. The curriculum description is essentially hierarchical. Each curriculum level has one or more teachers associated with it, one or more curriculum segments, an Intended Learning Outcome (ILO), and the next level, if there is one (left side of the figure). The center and right side of the figure show how the curriculum structure can be edited by adding, deleting and modifying individual curriculum levels and permit relating specific vocabulary and grammatical information with each segment in the curriculum level. A set

of more detailed interfaces (not shown) allow precise control of different aspects of the curriculum: segments and their links, grammar notes, and morpho-syntactic (MS) features. The latter encode grammatical information targeted by different curriculum levels. Currently MS-features pertain primarily to words but we will be extending them to cover multi-word fragments encoding specific grammatical constructs. The MS features currently in the system reflect Arabic grammar, but the framework is general and applicable to any language of interest.

2.2. Teacher Functionality

The functionality available to teachers pertains to two distinct modes: regular and testing. Here we focus on the former, leaving the discussion of testing to Section 4 below. The main functionality required by teachers in non-testing mode includes: 1) managing student accounts (accepting students into the levels they manage or changing the students' current level), 2) managing instructional materials, and 3) managing the dictionary. We restrict the discussion to management of materials and the dictionary.

The materials management interface allows teacher to view, add, delete and modify materials associated with the curriculum level of which they are in charge. Materials are characterized by the following attributes:

- **Level:** the level in which the material will be presented to its students;
- **Title:** materials title;
- **Type:** *Text, Audio* or *Video* (audio and video materials will be included in future versions);
- **Date:** the date when the material was created;
- **Role:** *Mandatory* (an obligatory part of the curriculum) or *Supplementary* (can be used for extra practice);
- **Description:** a brief description of content.

Materials are never completely deleted from the system. Modifications of materials result in a new copy being created, with the previous version having its status set to 'modified' and the new version having status 'active'. The reason for this 'lazy' approach to deletion is that tests and other information contained in the system may refer to older materials and eliminating them altogether from the system would create dangling references and loss of information. The materials management interface is shown in Figure 3.

The dictionary management interface for teachers (Figure 4) is a superset of the dictionary lookup interface also accessible by student users. It allows teachers to add, delete and modify dictionary contents in addition to searching for words and related information as students are allowed to do.

The upper and right-hand portion of the dictionary tool is devoted to search. There are two ways of looking up a word: *Exact match search and Incremental search*. In order to perform an exact match search, the user must uncheck the Incremental search checkbox, type in the

word and click on the Search button (the magnifying glass icon). The search can be done using or ignoring diacritics. For an incremental search—shown in the figure—the system keeps updating the list of suggestions under the search box until a final list is presented. There is no need to click on the search button. The user can then select a word out of the list to view details for.

Figure 3: Materials management interface

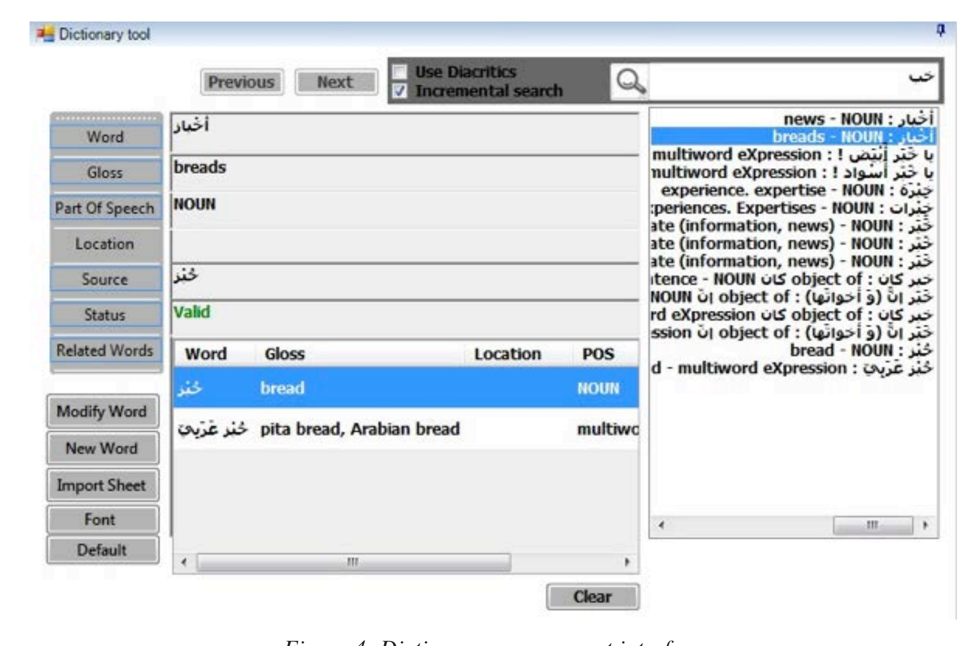


Figure 4: Dictionary management interface

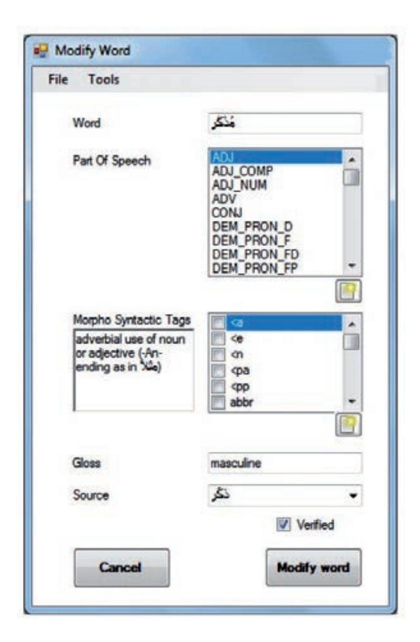


Figure 5 : Dictionary management interface

The upper part of the dictionary tool displays, upon request, different information about a given word: the word itself, which may be a lemma or an inflected form (depending on the dictionary contents), the gloss, the POS, the curriculum location (if specified), and related words. *The Source* field is a pointer to the nearest linguistic element from which the word is derived (in Figure 4, the plural has as source the singular). *RelatedWords* also include possible multi-word expressions in which the word occurs.

The lower left part of the dictionary tool is for modifying the dictionary. Individual words can be added or modified via the interface shown in Figure 5. Further information can be associated with vocabulary items including POS, different morpho-syntactic tags, the gloss, the Source (as described above) and the status of the word itself.

2.3. Student Functionality

The student interface is designed to give students access to a number of different functions that support the reading and learning process. The different tools supporting these functions appear in a standard arrangement but can be rearranged or minimized by using the standard functionality provided by the C# Dock Container interface to suit individual preferences.

Upon login, students are able to access materials suitable to their level. These materials are viewed through the text material interface shown in Figure 6. The text font characteristics can be changed, if desired, to make reading easier. The user can select a word and lookup information about it in the dictionary lookup tool. Note that student use of the dictionary tool is limited to looking up words and associated information and that options currently grayed out (diacritics, grammar tree, annotations) are currently placeholders for future additions.

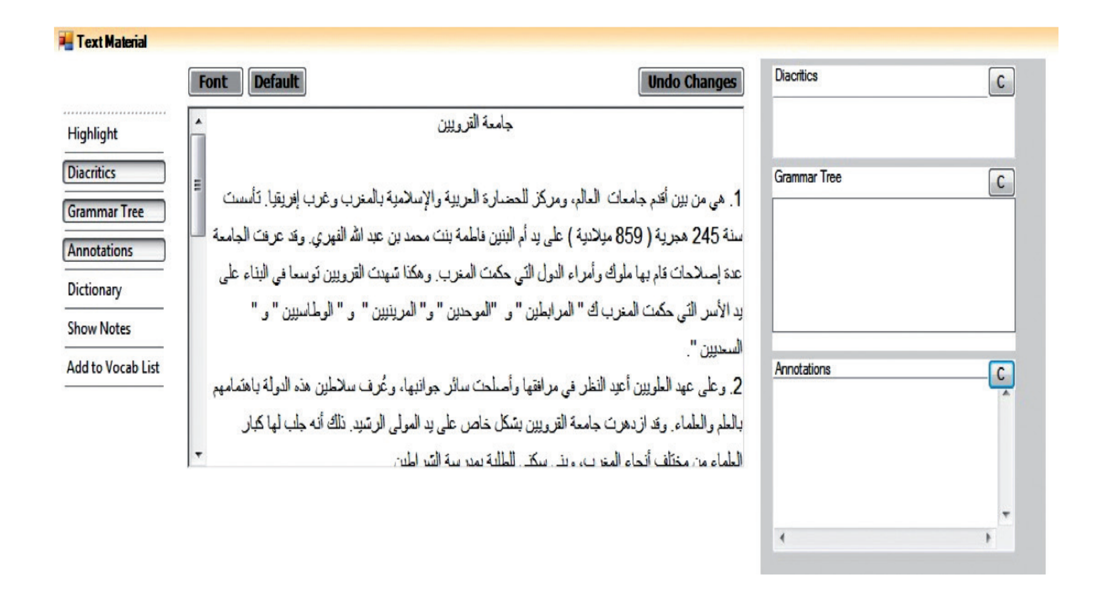


Figure 6: Text material interface

In addition to text display and dictionary lookup, the student interface provides two study tools: Student Notes and Student Vocabulary list. Students can create their own study notes about the materials they have read and manage those notes in different ways.

Student can view past notes they have created, filter them by subject, related text, related level or date create. They can also create a new note or delete an existing one and search the subjects of their notes in the search area. Whenever a note is added, the system automatically assigns it to the text currently open and the student's current level. Notes are viewed in two modes: either notes related to the current text or all the notes ever recorded.

The vocabulary list interface (Figure 7) is meant to help students have quick access to vocabulary linked to the level they are currently in or to build their own list of words to be reviewed whenever needed. At any time a student can add a new word to his/her list by writing it on in the text box and clicking on the plus button. This will show a list of corresponding words (in case of multiple results), from which the student should select one and add it to his/her own list. The user can also delete words from his/her list of words but not from the list linked to the level. For both views, the words are displayed in the form of an expandable tree; clicking on a word displays its meaning (Gloss) and Part Of Speech in the dedicated fields. These fields can be hidden or shown by the student whenever he/she wants to test his/her knowledge level. By clicking on the sort button, lists will be sorted alphabetically.

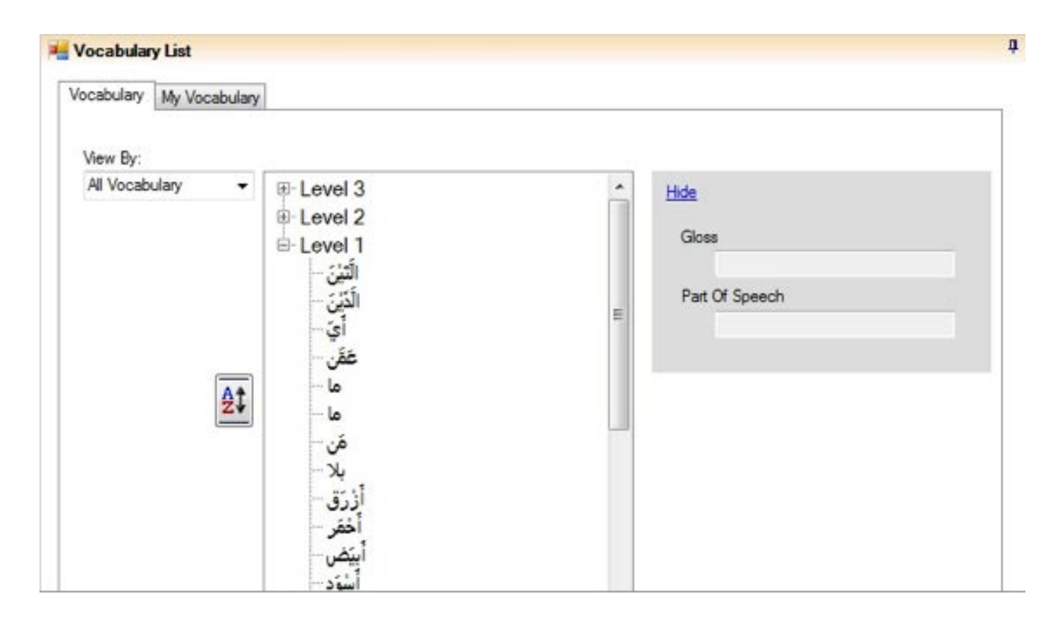


Figure 7: Vocabulary list interface

Last but not least, the Grammar Notes interface allows the student to get a list of HTML links to chapters and lessons of grammar grouped by level or by lesson. When clicking on the annotation of a specific word, from the text reading interface, the user can get the HTML links related to the word selected. For example, when clicking on the annotation of the word "book" which gives NOUN, the user get a list of HTML links with lessons and chapters for the noun structures. The default display of the list of links corresponds to grammar notes assigned by the admin to the student's actual level. A search option is provided to look for links that meet the student's needs. This facility is still under development. While the Grammar Notes already exist and can be normally browsed in a browser, the connection between them and the student interface is also still under development.

3. Testing Tools

The foregoing section provided an overview of the main components of the interface that have been developed, but omitted discussion of the testing (or assessment) functionality of the system. This functionality has already been prototyped but is still undergoing integration into the platform.

The testing component of the system consists of 4 modules: 1) *a test creation* module, used by the teacher to create tests (see Figure 8); 2) *a test taking* module, used by the student to take a test; 3) a *teacher test reviewing* module used by the teacher to examine student performance on the test; and 4) a *student test reviewing* module that allows students to examine their own performance on the test and the teachers' feedback. Questions that could not be automatically scored will be scored by the teacher, who can also provide further feedback.

There are essentially two types of tests: *material tests* and *level tests*. The former are intended for formative assessment, in order for the student and teacher to check strengths and weaknesses and inform further study. The latter serve to determine whether a student has achieved sufficient mastery of the concepts targeted by a curriculum level and is ready to go on to the next level. Apart from the difference in use, the tests are identically structured.

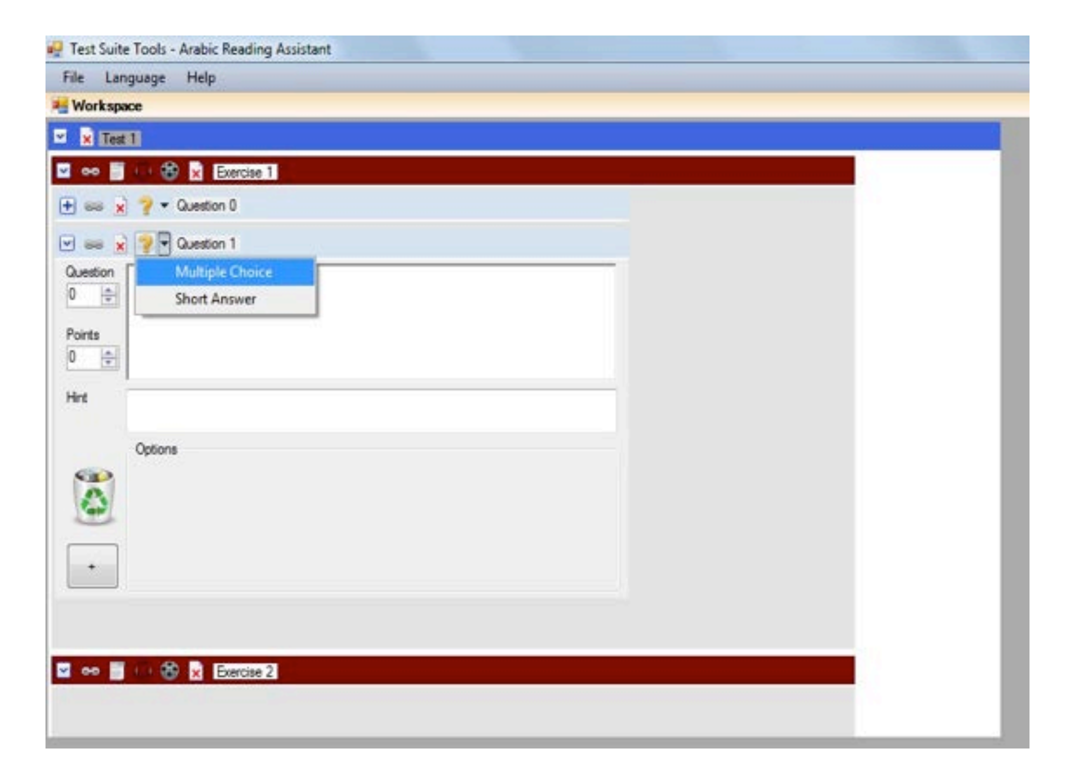


Figure 8: Test creation interface

The guiding philosophy behind the design of the testing modules, and particularly test creation, is flexibility and reusability. A test consists of a collection of *exercises*, and an exercise is a collection of questions. An exercise can be of two types: 1) material-specific, in which case all questions in the exercise pertain to the same (reading) material; or 2) general, in which case the exercise is just a container for a collection of questions, possibly but not necessarily targeting a specific skill. Questions themselves can be of three forms: 1) single-answer multiple-choice questions; 2) multiple-answer multiple-choice questions; and 3) short-answer questions. These three are sufficient to cover most common types of questions found in e-learning environments.

Multiple-choice questions can be automatically graded because the teacher can provide the correct answer(s) when creating the test. Short-answer questions can be semi-automatically graded if they have answers that are likely to fall in a predetermined set; answers that fall

outside that set will need to be scored by a teacher. In addition to providing answers or partial answers, test creators can also provide different weights (points) for each question and feedback associated with correct and incorrect answers. Given this structure, it is possible to manually or semi-automatically create tests by copying and modifying existing tests, swapping in exercises and questions that have not been used with a specific student before or that have been used but where the student has not performed correctly.

4. Representing and Entering the Language Curriculum

A centerpiece of the ARA platform is the data vocabulary, morpho-syntactic features and curriculum data that is loaded into the database and connects several components of the platform. Externally this data is stored in an Excel sheet, a small fragment of which is shown in Figure 9. (The current data is taken from part of the Al-Kitaab text book series).

VOCABULARY ITEM	LEMMA-ID	CURRICULUM LOCATION	POS	MS TAGS	GLOSS	SOURCE	ARG1	ARG1 ROLE	ARG2	ARG2 ROLE
أَقَلّ	>aqal~_2	K1.16:l	ADJ_COMP	m,<e	less than	أَقَلّ	مِن	than s.t.		
آخِر	|xir_1	K1.12:lg	ADJ_COMP	m,<e	the last …	آخِر	{ إسْم مُفْرَد }	of s.t., of s.o.		
أَحْسَن	>aHosan_2	K1.05:l	ADJ_COMP	m,<e	the best …	حَسَن	{ إسْم مُفْرَد }	of s.t., of s.o.		
أَحْسَن	>aHosan_2	K1.07:a	ADJ_COMP	m,<e	the best …	حَسَن	{ إسْم مُفْرَد }	of s.t., of s.o.		
أَصْعَب	>aSoEab_1	K1.12:l	ADJ_COMP	m,<e	the most difficult …	صَعْب	{ إسْم مُفْرَد }	of s.t.		
أَكْبَر	>akobar_2	K1.07:l	ADJ_COMP	m,<e	the biggest/oldest	أَكْبَر	{ إسْم مُفْرَد }	of s.t., of s.o.		
آخِر	|xir_1	K1.12:lg	ADJ_COMP	m,<e	last	آخِر				
أَواخِر	|xir_1	K1.12:d	ADJ_COMP	m,<e,bp	last	آخِر				
أَدْنَى	>adonaY_3	K1.02:f	ADJ_COMP	m,<e	nearer/est	دَنِي				
أَشْهَر	>a$ohar_2	K1.16:ao	ADJ_COMP	m,<e	more/most famous	شَهِير				

Figure 9: External storage of vocabulary data

Each row of the table shows: 1) an Arabic word with partial diacritics; 2) where it is found in the curriculum (volume and chapter) and how it is presented (e.g., as an item to be learned, in a grammar section, etc.); 3) the Part of Speech of the word; 4) some morpho-syntactic tags; 5) a gloss; 6) the word from which it is immediately derived (the source). Some entries have additional information regarding arguments used with the word. The vocabulary data in the Excel sheets, with which the database is populated, is one of the sources of information used to structure the curriculum. In combination with a description of where morpho-syntactic and other grammatical concepts are introduced, it defines the instructional curriculum, that is, the language concepts targeted at each stage. The curriculum itself can be structured, as shown earlier, through the ARA platform's administrator interface. Features describing morphology and grammar can be entered through another component of the administrator's interface, and individual vocabulary data can be entered through the dictionary interface. However, the Excel sheets provide a faster way of entering a large amount of vocabulary data.

The goal of keeping track in detail of morpho-syntactic features of specific words, and the curriculum location where these are introduced, is to support modeling the language concepts that a student should be familiar with at any given stage his/her learning trajectory. This in turn can be used to assess the student's progress and search for materials to provide additional

practice in areas where the student is weak, while avoiding repetition of concepts where the student has demonstrated sufficient mastery. It requires access to materials described in terms of the concepts on which the student is being tested, which in our project are tagged using the tools described in (Cavalli-Sforza *et al*., 2011).

5. Conclusions and Future Work

We have presented the ARA platform, conceived as a collection of integrated tools, to support the development and delivery of language curricula with a focus on reading and related activities. The platform includes interfaces targeted at three types of users, namely students, teachers, and an administrator who also plays the role of domain expert, master teacher, and instruction coordinator. At present, the components of the interface allows teachers to add instructional materials and supporting vocabulary, develop testing materials and review student performance, as well as manage student access to different courses. For students, the platform offers access to materials, the dictionary, personal study notes and simple vocabulary reviewing facilities, as well as access to external grammar notes. Testing materials are designed to be reused and composed flexibly for different types of testing at different times. Future work on the platform's interfaces will include completing integration of certain components, converting to a web-based interface, and porting to a mobile device, with a focus on Android-based systems. Linking to Google Translate, will provide approximate translations of texts (the quality of translation for Arabic is still not very high) and exposure to the sound of Arabic through the reasonably good text-to-speech system provided therein.

Some of the functionality present in ARA is undoubtedly useful in any learning environment; as such, it does not deviate substantially from that offered by many e-learning platforms. The special contribution of the ARA platform is rather the coordination of these elements through a complex database system which is designed to link vocabulary and grammar concepts targeted by the curriculum, to instructional materials and student performance in tests. The longer term goal, and the subject of current research, is to use this information to perform adaptive retrieval of materials thus offering students reading practice opportunities that focus on concepts targeted by the curriculum on one side, and individual strengths, weakness and interests on the other. Preliminary studies on the relationship between text features and reading level have already been carried out by ourselves and others (e.g., (Cavalli-Sforza & El Mezouar, 2014; Al-Khalifa & Al-Ajlan, 2010; Al-Khalifa *et al*., 2014)) and further work is in progress.

References

Al-Khalifa H. S., Al-Ajlan A. A. (2010). Automatic readability measurements of the arabic text: An exploratory study. *The Arabian Journal for Science and Engineering*, 35(2C):103-124.

Al-Tamimi A-K., Jaradat M., Aljarrah N., Ghanem S. (2014). AARI: Automatic Arabic readability index. *The International Arab Journal of Information Technology*, Volume 11, Number 4, July 2014.

Brown J., Eskenazi M. (2004). Retrieval of authentic documents for reader-specific lexical

practice. *Proceedings of InSTIL/ICALL Symposium*, 2004. (See reap.cs.cmu.edu/#Papers for more recent work.)

Brustad K., Al-Batal M., Al-Tonsi A. (2004 Part One, 2nd edition; 2007 Part Two, 2nd edition; 2001 Part 3). Al-KitaabfiiTacallum al-cArabiyya, A textbook for beginning Arabic: Part One, 2nd Edition, Georgetown University Press.

Cavalli-Sforza V., El Mezouar M. (2014). Characterizing the evolution of Arabic learners' texts: A mostly lexical perspective. Revue Asinag 9, Institut Royal de la Culture Amazighe, Rabat, Morocco, pp. 105-118.

Cavalli-Sforza V., Rehioui H., Bahri L. (2011). Tool for annotating texts with morphological and syntactic information. *4ème atelier international sur l'Amazighe et les Nouvelles Technologies sous le thème: Les Ressources Langagières: Construction et Exploitation*, Rabat, Morocco, pp. 249-259.

Krashen S. D. (2004).The Power of Reading: Insights from the Research, Heinemann.

Krashen S. D. (2004). Free voluntary reading: New research, applications, and controversies. *Proceedings of the RELC conference*, Singapore, April, 2004.

Krashen S. (2013). Reading and vocabulary acquisition: supporting evidence and some objections. *Iranian Journal of Language Teaching Research* 1(1):27-43.

Maamouri M., Zaghouani W., Cavalli-Sforza V., Graff D., Ciul M. (2012). Developing ARET: An NLP-based educational tool set for Arabic. *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT-2012)*, Montreal, pp. 130.

Nerbonne J., Smit P. (1996). GLOSSER-RuG: in Support of Reading. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996),* pp. 830-835.

Waring R., Nation P. (2004). Second language reading and incidental vocabulary learning. *Angleson the English-Speaking World*, Vol. 4.

L'apprentissage mobile de la langue amazighe : La situation d'apprentissage

Khadija Achtaich[1] **Habib Benlahmar[2]** **NaceurAchtaich[1]**
[1] Laboratoire d'Analyse, de Modélisation et de Simulation, LAMS,
Faculté des Sciences Ben M'Sik,
Université Hassan II, Mohemmadia, Casablanca
{k.achtaich, nachtaich}@gmail.com
[2] Laboratoire Technologies D'informationet De Modélisation, TIM,
Faculté des Sciences Ben M'Sik,
Université Hassan II, Mohemmadia, Casablanca
h.benlahmer@gmail.com

Résumé

Depuis 2001, l'enseignement de la langue amazighe est un chantier ouvert pour tout un chacun voulant y contribuer. Toutefois, aujourd'hui, il faut faire un effort colossal et avoir une volonté d'acier pour réussir à se frayer un chemin dans ce pactole de documents existant.

Nous nous sommes basés sur une idée simple, utilisant la puissance du terminal mobile, ne pas attendre que l'apprenant vienne solliciter l'apprentissage, mais venir à lui, en lui offrant le contenu propice selon son profil et son contexte spatio-temporel avec le concept de situation.

1. Introduction

Les Nouvelles Technologies d'Information et de Communication, notamment mobiles ont révolutionné l'apprentissage en le sortant de la salle de cours. Mine de rien, l'apprentissage renaît, sort de son bagne, quitte l'étroitesse de la salle de cours. Ce sont dès lors de nouveaux horizons qui s'ouvrent à l'apprentissage et donc à l'apprenant.

L'apprentissage situationnel que nous avons adopté en crédo, est celui qui se déroule dans un contexte authentique et un environnement culturel. Un paramètre est trivialement détectable est celui de la situation qui représentera dès lors l'anneau central de nos recherches.

Dans le domaine d'apprentissage, une place est particulièrement convoitée, c'est celle de l'apprentissage des langues qui n'ont jamais été aussi prisées. Une langue a particulièrement besoin de surfer sur cette vague technologique : la langue amazighe.

Cet article a donc pour but d'expliquer ce qu'est une situation et comment l'apprenant est mis au centre de cette dernière pour lui permettre un apprentissage optimal de la langue amazighe.

2. Concepts clés

2.1. Le terminal mobile

Du haut de leurs vingt ans, les dispositifs mobiles sont aujourd'hui à l'apogée de leur utilisation (Une pénétration d'à peu près 16% au Maroc)[1], les technologies de communication, du traitement d'information et du réseau sont plus développées que jamais. Tout individu est maintenant en mesure d'accéder à toute ressource informationnelle, à tout moment et à tout endroit, pour peu qu'il ait accès à un réseau sans fil. Le dispositif mobile se voit donc doté de la caractéristique de l'ubiquité. Cette singularité du dispositif mobile apparait dès lors comme une caractéristique nouvelle qui promet une révolution dans des domaines jusqu'alors confinés dans un cadre spatio-temporel délimité.

Plusieurs terminaux peuvent être réunis sous le vocable de dispositifs mobiles : l'agenda électronique personnel (PDA), le téléphone mobile classique, le terminal dédié, le téléphone intelligent (Smartphone) le cadet de la famille, et la tablette. Autrefois, seule les appels téléphoniques et les SMS combattaient l'isolement de ces dispositifs qui ne disposaient d'aucune connectivité. Les modes de connectivité se sont progressivement répandues pour donner vie à des standards (GSM, GPRS, EDGE, UMTS, LTE). Aujourd'hui les débits de transfert et la couverture géographique sont suffisants pour permettre des communications synchrones et instantanées ce qui permet d'intégrer de plus en plus de fonctions sur un seul terminal intelligent tel que le mailing, le chat, la vidéo conférence, la géolocalisation, etc. Les dispositifs mobiles n'ont plus rien à envier à leur prédécesseurs, les ordinateurs puisqu'ils sont aussi doté de leur propre système d'exploitation. De l'Android à l'IOS, en passant par RIM, Symbian ou encore Microsoft Windows Mobile etc., ils mènent tous une course technologique farouche accompagnée d'un développement moins méthodique des technologies de réseaux. Pour les tablettes par exemple, en 2013, le classement provisoire place l'Android confortablement en haut du podium avec 61,9% des parts de marché, suivi de son adversaire traditionnel, l'IOS avec 36,0%, le Microsoft arrive en troisième place avec 2,1%[2].

---

1 Statistique non officielle.

2 Gartner : Statistique du mois de février 2014 (vu sur http://www.gartner.com/newsroom/id/2674215 le 12 avril 2014)

		Galaxy S4[3]	iPhone5[4]	Nokia Lumia 1520[5]
Design	Système d'exploitation	Android	iOs	Windows
	Dimensions	136.6 x 69.8 x 7.9 mm	123.8 x 58.6 x 7.6 mm	162,8 x 85,4 x 8,7 mm
	Résolution	1920 x 1080 pixels	1136 x 640 pixels	1920 x 1080 pixels
	Ecran	5"	4"	6"
	Poids	130 g	112 g	206 g
Hardware	Connecteurs	- Micro-USB - Connecteur audio universel - MHL	- Connecteur Lightning - Connecteur audio universel	- Micro-USB - Connecteur audio universel
	RAM	2Go	1Go	2Go
	Mémoire de stockage	16Go/32Go/64Go	16Go/32Go/64Go	16Go/32Go
	Extension mémoire	MicroSD jusqu'à 64Go	NA	MicroSD jusqu'à 64Go
	Processeur	Cortex-A15 Or Cortex-A7	A7 and M7	Snapdragon™ 800
	Type du processeur	Octa-Core 1.6GHz Quad + 1.2GHz	Dual-Core 1.3GHz	Quad-core 2.2GHz
Batterie	Capacité de la batterie	2600 mAh	1570 mAh	3400 mAh
	Autonomie en standby	370 h	250 h	768 h
	Autonomie en communication.	17 h	10 h	20 h
Connectivité	4G LTE[6]	Oui	Oui	Oui
	NFC[7]	Oui	Oui	Oui
	WLAN	IEEE 802.11 a/b/g/n/ac HT 80	IEEE 802.11 a/b/g/n (2,4 and 5 GHz)	IEEE 802.11 a/b/g/n/ac
	Bluetooth	Bluetooth 4.0	Bluetooth 4.0	Bluetooth 4.0

3 Site officiel de Samsung, http://www.samsung.com/global/microsite/galaxys4/, vu le 12 avril 2014.
4 Site officiel d'Apple https://www.apple.com/iphone/compare/, vu le 12 avril 2014
5 Site official de Nokia, http://www.nokia.com/us-en/phones/phone/lumia1520/specifications/, vu le 12 avril 2014
6 Pour les opérateurs offrant les services de la 4G
7 Near-Field Communication

Photo	Résolution appareil principal	13 Mégapixels	8 Mégapixels	20 Mégapixels
	Résolution appareil secondaire	2 Mégapixels	1,2 Mégapixels	1,2 Mégapixels
	Applications	Google Play	App Store	Windows Phone Store
	Disponibilité	14 Mars 2013	20 Septembre 2013	Coming Soon
	Prix	5890 Dh[8]	5349 Dh[9]	NA

Tableau 1: Comparatif technique des leaders des Smartphones

2.2. L'apprentissage mobile

L'apprentissage mobile, appelé en anglais « M-learning » ou « Mobile Learning », est un apprentissage encore très peu utilisé dans les processus d'apprentissage et même encore très méconnu du grand public. Plus qu'une envie d'exiger plus des technologies existence, l'apprentissage mobile est un besoin assouvi par le potentiel technologique des dispositifs mobiles tels que les PDA, les Smartphones, les Tablettes, etc.

Bien que présente dans l'apprentissage électronique, l'apprentissage mobile encre encore plus la distance à l'apprentissage et introduit le concept de mobilité, jusque-là absente dans toutes les formes d'apprentissage que nous qualifions de classiques ou traditionnelles. La notion ubiquitaire ou de l'omniprésence est intrinsèquement introduite avec la mobilité.

2.3. La langue amazighe

La langue amazighe est parlée dans plusieurs pays d'Afrique, notamment le Maroc, l'Algérie, la Tunisie, la Lybie, l'Egypte, la Mauritanie, le Mali et le Niger.

Au Maroc, la langue amazighe est légitime depuis le 17 Octobre 2001 et est officielle depuis juillet 2011. L'Institut Royal de la Culture AMazighe (IRCAM) est chargé de la promotion de la culture et le développement de la langue amazighe. En collaboration avec les autorités gouvernementales et les institutions concernées, l'IRCAM concourt à la mise en œuvre des politiques devant permettre l'introduction de l'amazighe dans le système éducatif et assurer son rayonnement dans l'espace social, culturel et médiatique, national, régional et local. Depuis juillet 2002, le Tifinaghe est adopté pour l'écriture de la langue amazighe et ce n'est qu'en juin 2004 que le Tifinaghe est reconnu à l'échelle internationale par l'organisation internationale ISO-unicode.

L'enseignement de la langue amazighe en revanche reste bien au-dessous des espérances. Il se base aujourd'hui principalement sur quelques manuels et documents fournis par l'IRCAM

8 Prix chez Maroc Telecom en abonnement de 24 mois.
9 Prix chez Meditel en abonnement de 24 mois.

ou par le Haut-Commissariat à l'Amazighité en Algérie (HCA). Ces documents sont souvent disponibles en format papier (consultable aux bibliothèques spécialisées ou sous forme de publications) ou en format électroniques (CD-ROM, PDF téléchargés sur Internet ou cours interactifs pour les niveaux primaires).

2.4. L'apprentissage mobile de la langue amazighe

Plus qu'une envie de développer une nième plateforme d'apprentissage de langue, c'est un besoin d'offrir un contenu d'apprentissage aux personnes désirant apprendre la langue amazighe. Nous œuvrant dans le développement d'une plateforme qui peut supporter toutes les langues pour lui offrir plus d'horizons et plus d'ouverture.

Nous avons par ailleurs essayé de mimer l'apprentissage d'un enfant en bas âge. En effet un bébé de deux ans aurai appris en ces deux années de sa vie plus qu'il n'apprendra dans le reste de sa vie. Son apprentissage de la langue passe par l'apprentissage de lexies et non par des phrases complètes. Un vocabulaire riche permettra de gagner en confiance et « d'oser » communiquer par le billet de cette langue.

3. Vers une situation d'apprentissage

3.1. Définition d'une situation d'apprentissage

Plusieurs définitions sont proposées dans la littérature dont la plupart relèvent l'implication de l'enseignant et de l'apprenant dans la situation d'apprentissage, d'autres font sa corrélation avec le savoir ou encore la connaissance. Nous nous proposons de présenter notre propre définition :

Une situation d'apprentissage est l'union, de l'environnement spatio-temporel, et du profil de l'apprenant ainsi que du contexte technologique et des ressources numériques, passible d'amener l'apprenant vers la connaissance.

De ce fait l'apprenant est au centre de son processus d'apprentissage mais de façon très subtile, comme indiqué sur le schéma suivant :

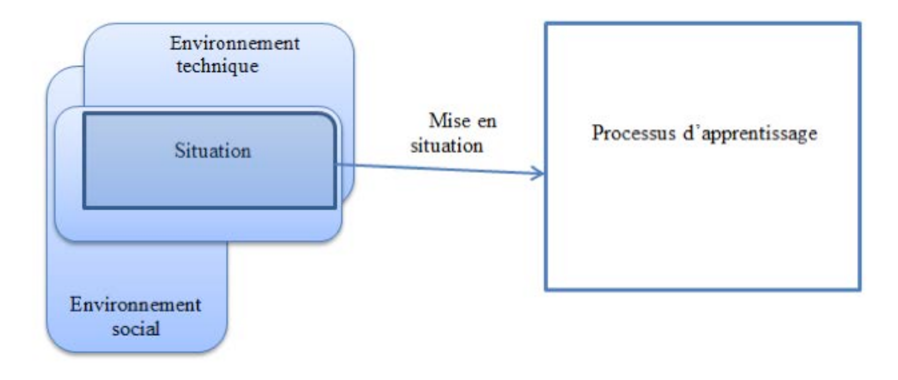


Figure1 : Situation d'apprentissage

De ce schéma deux axes principaux conduisant à deux valves de recherche indépendantes :

- L’interaction de la situation originale avec l’environnement technique et social.
- Les critères de mise en situation

Dans la suite de cet article, nous discuterons brièvement ces deux volets dans l’optique de détailler chacun d’eux dans un article indépendant.

3.2. L’interaction avec l’environnement

3.2.1. L’environnement technique

L’apprentissage se déroule via un terminal mobile, l’environnement technique regroupe toutes les caractéristiques du terminal mobile à travers lequel se passe l’apprentissage. Supposons qu’un même apprenant utilise tantôt son smartphone, tantôt sa tablette avec les quelques différences techniques suivantes :

Smartphone	Tablette
3G / 4G / Wifi	Wifi
Appareil principal et secondaire	Appareil principal
Taille de l’écran limitée ~5"	Taille assez étendue ~9"
Espace de stockage libre limité	Espace de stockage libre suffisant
…	…

De ce fait, une même situation déclenchée par le même apprenant, peut donner lieu à deux exécutions différentes.

3.2.1. L’environnement social

Un apprenant est affecté lors de son apprentissage par son environnement social. C’est ce cadre qui nous permet de définir le profil de l’apprenant. Le cadre ou l’environnement ou encore le contexte social est l’ensemble d’éléments et de conditions qui entourent l’individu constituant son cadre de vie et susceptible d’agir sur lui ou d’interagir avec lui. L’éducation, les origines, les études, l’âge, le genre, les relations, les conditions de vie, le travail sont tous des éléments constituants l’environnement social.

En effet, des situations peuvent être proposées à un groupe d’individus avec des conditions sociales particulières et ne peuvent être proposées à d’autres, le tableau 2 liste cinq profils sociaux-culturels. Le tableau 3 en revanches présente un exemple mettant en évidence cette affirmation.

Profil	Contexte social
Profil 1	23 ans, Masculin, Diplômé universitaire, Sportif, Fan de voyages, Sociable …
Profil 2	10 ans, Féminin, CM1, Pratique la dance classique, …
Profil 3	53 ans, Masculin, Père de famille, Responsable marketing, Sédentaire, Sort très peu …
Profil 4	38 ans, Féminin, Femme au foyer, Niveau secondaire, Fan de ménage et de couture …
Profil 5	45 ans, Féminin, Professeur Chercheur, Sociable, Enfants en bas âge …

Tableau 2: Exemple de profils

Situation	Profil 1	Profil 2	Profil 3	Profil 4	Profil 5
Stade de football	Oui	Non	Oui	Non	Non
Port	Oui	Non	Oui	Non	Non
Université	Oui	Non	Non	Non	Oui
Au marché	Non	Non	Oui	Oui	Non
Parc de jeux	Non	Oui	Non	Oui	Oui
Restaurant	Oui	Oui	Non	Oui	Oui
Ecole	Non	Oui	Oui	Oui	Oui

Tableau 3: Disponibilité des situations selon les profils

De même, une même situation peut présenter des contenus différents toujours selon les différences sociales présentées par les apprenants.

Reprenons une des situations précédemment présentée (Un restaurant par exemple), On peut éventuellement la proposer à tous les profils précédents mais avec des variantes comme indiqué sur le tableau 4:

Profil	Variantes
Profil 1	Le mobilier, le menu, les ustensiles …
Profil 2	Le nom de quelques plats, le mobilier élémentaire (Chaise, Table …), Le couvert …
Profil 3	Le mobilier, le menu détaillé, les produits utilisés en cuisine …

Profil 4	Le mobilier en détaille (Fauteuil Vs Canapé, Lustre Vs suspension …), Les ingrédients du menu détaillé
Profil 5	Le mobilier, Les types de dressages, l'art de la table …

Tableau 4: Variantes de situation

3.3. Les critères de mise en situation

La grande avancé du mobile en terme d'apprentissage c'est la possibilité d'offrir un apprentissage personnalisé aussi bien que toujours disponible. De ce fait nous divisons la mise en situation en deux types : Une mise en situation volontaire et une mise en situation spontané.

La mise en situation volontaire rejoint la forme classique de l'apprentissage électronique ou l'apprenant exprime le besoin d'apprendre et fait lui-même le choix du contenu. Ceci étant le besoin exprimé est influencé par le contexte où se trouve l'apprenant. Par exemple, un apprenant qui se trouve sur station de ski a plus de chance d'exprimer le besoin d'apprendre ce qu'est une luge, un télésiège ou un téléphérique. Par contre, nous ne devrons pas trop s'attendre à ce qu'il demande ce qu'est un éléphant ou ce qu'est une rivière.

Quant à la mise en situation spontané, est plus dans l'esprit de l'apprentissage mobile. L'idée ici est de ne pas attendre que l'apprenant exprime son besoin mais de lui proposer la situation la mieux adapté selon son contexte spatio-temporelle. Deux dimensions sont exploitable : la dimension temps et la dimension espace.

Par exemple, pour une situation baptisée « stade de football » serait intéressante en période de la coupe du monde même pour des apprenants ne manifestant pas d'intérêt particulier pour le football. Idem si un apprenant est à la plage, nous supposant que cela l'intéresserai d'apprendre d'avantage sur la plage.

4. Conclusion

Sur cet article, nous nous somme mit à la place de l'apprenant Marocain qui exprime la volonté d'apprendre la langue amazighe, mais qui pour une raison ou une autre risque de rapidement perdre la motivation. Une fois que l'apprenant manifeste son désir d'apprendre la langue, nous nous chargeons du reste en lui proposons des situations en harmonie avec lui, et avec « lui » nous entendons son environnement social et technique.

TICAM'14, p. 33-39

الأمازيغية والتعليم الرقمي
كفاية الحجاج نموذجا

بودريس بلعيد
أستاذ التعليم العالي– مركز تكوين مفتشي التعليم، الرباط

مقدمة

تشكل معالجة الوضعيات التعليمية بالمقاربة المركبة أحد مظاهر التربية الذي صاحب الثورة المعرفية. ويشكل التعليم الرقمي احد تجليات التفكير المركب، لأنه يدبر وضعية تعليمية–تعلمية تعتمد على التفاعل بين المتعلم والوثيقة الرقمية التي تتحمل كل تبعات التفاعل الممكن والفعال والمحقق للنتائج المخطط لها. وتحاول الدراسة تجسيد أحد أشكال هذا الإنجاز، من خلال التخطيط الرقمي لكفاية الحجاج في درس اللغة الأمازيغية.

وتتكون الدراسة من قسمين:

- قسم تمثله الوثيقة الرقمية التي تجعل من كفاية الحجاج موضوعا لها؛
- ويتمثل القسم الثاني في تقديم الحيثيات التي أدت إلى مختلف المرتكزات النظرية التي أسست القرارات التي صاغت الوثيقة. وهذا القسم الثاني هو موضوع هذه الورقة.

1. الأساس النظري والتفسيري للوثيقة

يستوجب تقديم الحجاج في درس اللغة الأمازيغية إنجاز التحليل الإبستيمولوجي للحجاج. ويتضمن هذا التحليل ما يلي:

1.1. الاستيعاب التاريخي لتطور دراسة الحجاج منذ أرسطو[1]، الذي أطر الحجاج كمحاولة للإقناع، في إطار البلاغة والخطابة، مميزا له الإفحام الذي يدخل في إطار الجدل. وقد كانت صرامة التفكير الديكارتي سببا في تواري الدراسات البلاغية–الخطابية، مما شكل السبب في قيام البعث البلاغي–الخطابي الذي أنجزه بيريلمان[2]. وإذا كانت أعمال هذا الأخير قد شكلت نقطة تحول جذرية في اتلول

1 Aristote, (trad. Pierre Chiron), *Rhétorique*, Paris, Flammarion, 2007,
2 Chaïm Perelman, Lucie Olbrechts-Tyteca, (1958) , Traité de l'argumentation, éditions de l'université de Bruxelles, 2008

الحجاج، فإن مساهمة تولمين[3] قد أضفت على الحجاج مسحة ديناميكية مهدت لاستيعاب الحجاج لمختلف ظواهر التواصل وتحليل الخطابات ذات الأطروحة، ومنها الإشهار والخطاب السياسي. وكان لتدخل البراغماتية، على المستوى الفلسفي واللساني[4] وظهور البراديغم المعرفي، بالغ الأثر في ربط الحجاج بنوعي التفكير النقدي والإبداعي.

2.1. التمييز في هذا الجرد بين الأساسي والثانوي، بالشكل الذي يتمكن فيه المخطط للتعلمات من إدراك الأسبقيات سواء من حيث الأهمية أو من حيث انعكاس ذلك على المكونات بحيث يكون اعتماد بعضها على بعض سلسا من الناحية المنطقية والديداكتيكية.

وإذا كان استيعاب كل هذا التاريخ من جهة، والتمكن من إنجاح النقل الديداكتيكي من جهة ثانية [5]، يشكلان المسار الأكثر ملائمة، فإن ذلك يحتاج ليس فقط إلى وقت استيعاب الموضوع ودقة النقل الديداكتيكي، بل يحتاج أيضا إلى وضوح المشروع الشامل الذي تندرج فيه كفاية من الكفايات التي يهدف درس الأمازيغية إلى تنظيمها أو تطويرها.

2. متن الدراسة

وعليه، فإن المقام الذي نحن بصدده لا يستوعب الحديث عن النقل الديداكتيكي بهذه الشمولية، الذي يضمن النسقية والفعالية، لذلك سنعتبر ما هو مثبت في الكتاب المدرسي الأساس الضمني لهذه المداخلة دون جعله موضوعا للمساءلة أو إعادة النظر فيه. أي أن ذلك سيعفينا من إنجاز النقل الديداكتيكي، بكل ما يتطلبه ذلك من تخطيط لتسلسل المعارف المرتبطة بالحجاج وتناسق هذا التسلسل وتفاعل عناصره وانتهاء بالتقويم. لأننا سنركز على الإمكانيات الرقمية وأحد أشكال تنظيم المقطع المعرفي وترتيبه. ونسمي المقطع المعرفي «المقدار المعلوم والمحدد بما يتأسس عليه من معارف وما يمهد له من أخرى في إطار سيرورة متنامية من بناء معارف الإطار المرجعي للمتعلم». والمقطع، بهذا المعنى وفي تقديرنا النظري، تركيز على معرفة محددة (limitée) مع الأخذ بعين الاعتبار بأنها لحظة في سيرورة، أي أنها تتأسس على ما سبقها من معارف وتمهد لما يليها.

3 Stephen E. Toulmin, (1958), the uses of arguments, Cambridge University Press, updated edition, 2003.
4 Jean-Claude Anscombre, Oswald Ducrot, L'Argumentation dans le langage Editions Mardaga, 1983.

5 انظر : بودريس بلعيد، استراتيجيات الحجاج وأساليب الاشتغال المعرفي عند تلاميذ التعليم الثانوي، بحث لنيل دكتوراه الدولة، غير منشور، كلية الآداب والعلوم الإنسانية، فاس، 2002.

وبخصوص المقطع المعرفي الذي يشكل مادة هذا العمل، فنحن نعطي الأسبقية للمعالجة الرقمية على حساب كل ما يمكن أن يتعلق بنجاعة المقطع المعرفي، الذي يمتد على مدى أربع سنوات. وهذا لا يعني أن نجاعة المقطع المعرفي لا تحتاج لمعيار النجاعة بل بالعكس لأن الشرط الأول هو هذه النجاعة، ولأنها بهذه الأهمية الرئيسة، فإننا لا ندخله في المجهود المبذول في هذه المداخلة. وذلك رغبة في التركيز على أشكال خلق التناسب والتلاؤم والفعالية بين المقطع المعرفي والإمكانيات الرقمية.

ويستدعي النقل الديداكتيكي اعتبار العناصر التالية:

1. **سلم التكييفات والقدرة على الملائمة:** تعتبر مسألة التكييفات التي تخضع لها المادة المعرفية من أهم المعالجات التي تخضع لها المادة الدراسية. وتأخذ هذه المسألة الملامح التالية:
2. **تحديد المقطع المعرفي:** يتعلق الأمر بتحديد الشحنة المعرفية المراد الاشتغال عليها ديداكتيكيا. ويحدد هذا المقطع، من الناحية الطبيعية، من البداية إلى تمكن المتعلم من آليات البحث والتكوين الذاتي. وبالنسبة للحجاج يمكن القول بأن البداية يمكن أن تتموقع في سن الثالثة، ويمكن مؤقتا تحديد الاشتغال بالمعنى المدرسي على الحجاج مع الحصول على الباكالوريا.
3. **ترتيب تطور الحجاج وتدرجه:** تتجلى الواجهة الثانية في التمييز بين الأساسي والتابع في بناء "كل" المادة المعرفية التي تهم الحجاج. ونعني بكل المادة ما يقع دون الدراسة المتخصصة في الحجاج، الذي يمكن أن يتجلى في البحث العلمي في شعب الفلسفة أو الآداب أو اللغات أو العلوم القانونية والسياسية. ونظرا للطابع الاستراتيجي لهذه المسألة، وحاجتها، بالإضافة إلى الانسجام، ضرورة توفرها على الانسجام مع المواد الحاملة للحجاج وباقي المواد المدرسية، فإننا نكتفي بالاشتغال على التدرج الذي نجده في (ⵜⵉⴼⴰⵡⵉⵏ ⴰ ⵜⴰⵎⴰⵣⵉⵖⵜ):

وعلى هذا الأساس، يتمثل متن الدراسة في حضور الحجاج من السنة الثانية من التعليم الابتدائي إلى السنة الخامسة، وتوضيح ذلك فيما يلي:

- السنة الثانية: التعرف على الحجج (مع) و(ضد)، ترتيب الحجج، التحكم في الروابط المنطقية؛
- السنة الثالثة: التعرف على الحجج (مع) و(ضد)، ترتيب الحجج، التحكم في استعمال الحجج؛
- السنة الرابعة: التعبير عن الاختيار وتبريره بالحجج وإنتاج حجج ضد وتنظيم الحجج؛
- السنة الخامسة: الإقناع بواسطة حجة السلطة وحجة المنفعة وتحديد الوضعية المناسبة لكل نوع من الحجج.

على أن يتجلى المتن بالملامح التالية :

- تحديد الهدف، وتوزيع الأنشطة بين الحجاج في الشفوي والحجاج الكتابي؛
- إدماج الهدف في السياق العام للمعرفة (التدرج في الابتدائي، التدرج الموالي في الإعدادي والتأهيلي)؛
- التحليل الفارقي: الذي يسمح بإدراك المتعلم وتمييزه بين الأهم من المهم والأساسي من الثانوي، بشكل يكون الانتقال متدرجا ومنطقيا؛
- وضوح الامتدادات والتمفصلات: ليس المقصود هو الإدماج المباشر للثنائيات التالية: الحجاج والسرد، الحجاج والإخبار، الحجاج والتفكير النقدي، الحجاج والتفكير الإبداعي، الحجاج وحل المشكلات واتخاذ القرار، الحجاج والأسلوبية، بل يتعلق الأمر بخلق بذور التفاعل بين هذه المكونات، بحيث يؤدي التقدم في التعلمات إلى التقدم التدريجي في بناء مختلف الكفايات بشكل تفاعلي وغير معزول.

3. الخصائص الرقمية للوثيقة الحجاجية

تمثل التكنولوجيا التربوية الحديثة حلا أصيلا للمسألة التواصلية. وتتمثل هذه الأصالة في بناء عملية تواصلية تتميز بالتالي:

1. **غياب إمكانية التعديل:** يشكل اللقاء بين الوثيقة الرقمية والمتعلم لحظة لا مكان فيها لاستدراك النقص أو غياب الانسجام، بشكل نهائي. إذ يتحولان إلى عوائق للاستعمال السليم وتنفير مؤقت من الوثيقة الرقمية. مما يؤدي إلى خلق عوامل الإحباط التي تؤدي إلى خسران الوثيقة الرقمية لقيمتها التعليمية والإسهام في النفور من استعمالها في تعلم اللغة الأمازيغية.
2. **غياب نية المرسل في التدخل:** تشكل نوايا وإرادة مصمم الوثيقة الرقمية، بعد الانتهاء منها، عوامل ومحددات دون أي حضور يذكر. أما من الناحية البعدية فإن حضورها يتراوح بين قطبين:

- إما أنها مفعلة بشكل عملي ومكتف بذاته ومأجرأ بما يمكن من الاستفادة منها في وضعية وثيقة– متعلم ؛
- أو أن هناك نقصا في التصور وتنفيذه إجرائيا، على أن كل نقص هو بمثابة خطوة مضادة للأهداف الأولية للوثيقة.

3. **الملل وخفوت التحفز:** يعتبر التقدم الملموس في التعلمات وكلفتها الدنيا من حيث المجهود من المحفزات التي تمنع الملل من التسرب إلى العلاقة بين الوثيقة والمتعلم، ذلك أن القرار الأخير في استعمالها والعمل على الاستفادة منها يعود للمتعلم، وذلك حتى في الحالة التي تدخل الوثيقة ضمن استراتيجية أعم يتحكم فيها المدرس.

4. **ضرورة بناء الوثيقة بالمعنى التام:** مما سبق تأتي أهمية بناء الوثيقة بأكثر ما يمكن من النظرة الشمولية التي تدمج كل الاعتبارات المعرفية والتقنية.

إن هذه الخصائص هي التي تجعل من الوثيقة الرقمية وثيقة توضح منطق اشتغالها وتعمل على إيصاله وتملكه من طرف المستعمل. ويمكن اختصار هذا المنطق في:

1. **وضوح قواعد الاشتغال:** تشتغل الوثيقة بقواعد واضحة تعبر عنها إيقونات ذات إيحاءات غير قابلة للالتباس كما في حالات الأسهم التي تبين الاتجاه؛

2. **محدودية القواعد:** تتحدد مجموع القواعد التي تشتغل بها الوثيقة

3. **انسجام مكونات النسق:** تشكل الوثيقة نسقا منسجما ومتماسكا قابل، من طرف المتعلم، للإدراك باعتباره كذلك؛

4. **سهولة الاستعمال ونجاعته:** إن كل صعوبة غير معالجة، كهدف تعلمي تعتبر إسهاما فعالا في إفشال تحقيق وثيقة لأهدافها.

بالإضافة إلى هذه الخصائص التقنية، هناك اعتبارات ديداكتيكية يجب توفرها في الوثيقة الرقمية. ويتعلق الأمر بالاعتبار الديناميكي لتفاعل العوامل التالية:

1. **استحضار المعارف السابقة:** تشكل المعارف السابقة مجموع مكتسبات المتعلم، بل تشكل أيضا عاملا مؤثرا في إدراك المعارف الجديدة وأشكال تخزينها؛

2. **احتساب التمثلات:** تعتبر التمثلات[6] أحد مداخل فهم اشتغال التفكير عند المتعلم. وعند معالجتها يتم تكييفها مع المعارف الجديدة أو يتم خلق الانسجام بينها وبين هذه الأخيرة في حالة وجودها كعائق معرفي. أما كيفية احتسابها فتتمثل في الاختبارات المتعددة الاختيارات التي تتأسس على طرح التمثلات المفترضة ضمن الأجوبة المتاحة؛

3. **تنويع مسارات تنظيم المعرفة:** وذلك لإتاحة الفرصة لكل من التعلمات الخطية والتعلمات المتشعبة، لخلق نوع من الحركية والتوافق مع ريتمات المتعلمين؛

4. **احتساب النشاط التفكيري للمتعلم:** تتجلى أهمية الانخراط في بناء التعلمات في الفرص التي تتيحها للمتعلم من أخذ زمام المبادرة في معالجة المعلومات وتنظيمها وتخزينها واستعمالها وإعادة استعمالها. أما كيفية احتسابها فتتجلى في الفراغات التي على المتعلم أن يملأها بنتائج تفكيره وتعقله للموضوع، سواء أثناء مسار التعلم أو أثناء التقويم.

6 Patrick Lemaire, Psychologie cognitive, édition de Boeck, Bruxelles, 1999.

4. الإمكانيات التربوية والتعليمية للوثيقة الرقمية

1. **التفريد والاستقلالية:** يشكل تفريد التعليم وتحركه وفق وتيرة خصوصيات التلميذ أحد المسارات التي تحقق التعليم الفعال بالمعنى الذي يحدده روجرس، أي التعليم الذي يعني شيئا أصيلا في سيرورة تعلمات المتعلم. كما أنه من ناحية أخرى، يعتبر مساعدا على خلق الاستقلالية بما تتضمنه من تعلم كيفيات التعلم والمعالجة الشخصية والإرادية للوضعيات–المشكلات إيجاد الحلول الإبداعية أثناء اتخاذ القرارات.

2. **استيعاب المنطق التكنولوجي:** ينتظم الاشتغال الرقمي بحساب مضبوط لا يقبل أي نوع الالتباس أو الانزياح. وهذا يعني من الناحية الإبيستيمولوجية، ربط علاقة وضوح ودقة وتنظيم عبر تلك التي يشهدها التفاعل بين الناس. لأن التفاعل مع الوثيقة الرقمية يتأسس على منطق شفاف، لا يلبث المتفاعل معه أن يدرك منطق الاشتغال وتلقي تأثيراته وانعكاس ذلك على التنظيم التدريجي لمنطق المتعلم.

3. **استيعاب التدبير الذاتي للتعلمات:** يؤول بناء المعارف بتنشيط المكتسبات السابقة للمتعلم إلى التمكن التدريجي لهذا الأخير من تدبير بناء تعلماته حسب الريتمات المتنوعة بتنوع معطيات المعرفة القبلية التي تتأسس عليها وتتفاعل معها التعلمات الجديدة.

4. **استيعاب الضبط والتخطيط من طرف الوسيط:** في حالة استعمال المدرس للوثيقة الرقمية

5. **تملك العمليات الأساسية:** تمكن الوثيقة الرقمية من الوجود الفعال في وضعية التقويم والتقويم الذاتي ذات الخصائص الإيجابية التالية:

- **التصحيح الفوري:** يؤدي التصحيح الفوري إلى تجنب أخذ الخطأ لموقعه في المرجعية المعرفية للمتعلم، والتمهيد للمعلومة الصحيحة بأن تجد لها المكان المناسب. وهذا يعني خلق فرص التقدم في الاكتساب بالمعارف المصححة والمفكر في أسباب أخطائها؛

- **التكرار المساعد على الاسترجاع الوظيفي:** تقدم الوثيقة تكرارا مقصودا للوحدات المعرفية أو للأسئلة، رغم اختلاف صياغاتها، مما يساعد الذاكرة القصيرة المدى على معالجة المعلومات وحسن تخزينها. وهذا ما يوفر الشروط الضرورية للاسترجاع المدعم بالاستعمال في وضعيات متعددة، مما يزيد من فرص التملك الفعلي للمعلومات.

- **تنويع الحوامل:** تتنوع الحوامل بين لفظية–إيقونية وبين سمعية/بصرية/سمعية–بصرية مما يكسب التفكير مرونة وقدرة على التمييز بين الأساسي والثانوي مما يسهل الفهم والنقد والإبداع.
- **انعكاسات المنطق الرقمي:** يمكن المنطق الذي تشتغل به الوثيقة الرقمية من تنمية مهارة التخطيط عند المتعلم. كما أن الاستقلال في تدبير اللحظة التعلمية يساعد السيرورات الميتامعرفية على الاشتغال. كما أن تنشيط آليات التقويم والتقييم التشخيصي ينمي كفايات النقد والإبداع.

خاتمة

من خلال ما سبق، يتبين أن المهام الأساسية للوثيقة الرقمية تتجلى، على الأقل، في ما يلي:

1. تملك الشحنة المعرفية وفهمها: ويتم ذلك من خلال تنويع وضعيات الحجاج بالنسبة لنفس الشحنة المعرفية، مما يجنب النمطية والسطحية واللفظية في الاكتساب؛
2. تطبيق الشحنة وتحليلها وتركيبها: يشكل التطبيق والتحليل والتركيب مؤشرات دالة على تملك المعرفة واستعمالها خارج السياق الملموس لوضعية التعلم والاكتساب؛
3. ممارسة النقد والإبداع: المتعلق بممارسة تفنيد الحجاج المقدم والعمل على تجاوزه، ثم العمل على إبداع سياق حجاجي.

La connaissance et la perception des jeux vidéo comme outil didactique par les enseignants stagiaires de la langue Amazighe

Ibrahim Ouahbi[1] Hassane Darhmaoui[2] Fatiha Kaddari[1]
Abdelrhani Elachqar[1] Mohamed Regragui[3] Soufiane Lahmine[1]
[1] LADIPEC, Faculté des Sciences Dhar Almahraz, Université sidi Mohamed ben Abdellah, Fès, Maroc
ibrahim.ouahbi@usmba.ac.ma, kaddari@yahoo.fr,
abdelrhani.elachqar@usmba.ac.ma, lahmine.soufiane@gmail.com
[2] Center for Learning Technologies, Al Akhawayn University, Ifrane, Maroc
h.darhmaoui@aui.ma
[3] CRMEF, annexe régionale de Nador, Maroc
reg.mohamed@gmail.com

Résumé

Cette étude vise l'exploration des attitudes des futurs enseignants de la langue Amazighe envers les apprentissages par les jeux vidéo. Notre objectif est de promouvoir l'innovation pédagogique dans l'enseignement de la langue amazighe au primaire à travers le "Serious Game". Pour ce, une vingtaine d'enseignants stagiaires ont été appelé à remplir notre questionnaire sur leur perception envers ce mode d'enseignement. Ce questionnaire s'articule autour de 3 axes: leur expérience avec les jeux vidéo, leur attitude envers les jeux vidéo en général et leur attitude envers l'intégration des jeux vidéo en éducation. L'analyse des résultats a montré que la plupart des futurs enseignants jouent, ou bien, ont déjà joué aux jeux vidéo. Leur perception envers l'utilisation des jeux vidéo dans la classe diffère selon l'expérience antécédente de chaque stagiaire avec les jeux. Après cette pré-enquête, nous avons mené des activités de veille et de sensibilisation sur le "serious game". A la fin de ces séances un autre questionnaire a été distribué aux mêmes enseignants stagiaires pour mesurer l'évolution de leur perception envers les jeux vidéo supportant l'apprentissage. L'analyse des résultats a montré une bonne prise de conscience sur les potentialités des jeux sérieux en classe et sur les différentes possibilités d'exploiter les jeux vidéo pour faciliter l'apprentissage de la langue amazighe. Nous envisageons par la suite élargir et planifier notre expérimentation avec un public plus large d'étudiants professeurs de différentes disciplines.

1. Introduction

L'industrie du jeu vidéo dépasse actuellement l'industrie du film en termes de profits[1]. En Septembre 2013, par exemple, le jeu "Grand Theft Auto V", est devenu la première pièce de divertissement y compris les films et albums. Il a atteint 800 millions de dollars de

1 http://blogues.radio-canada.ca/triplex/2010/11/15/lindustrie-du-jeu-video-plus-lucrative-que-celle-du-cinema/; consulté le 29/05/2014.

ventes mondiales durant les premières 24 heures après son lancement. En seulement trois jours, il a franchi la barre d'un milliard de dollars. A titre de comparaison, le film Avatar, la superproduction hollywoodienne, a pris 17 jours pour atteindre ce seuil. Il est à noter que des jeux comme "AngryBirds" et ses diverses incarnations ont été téléchargés plus d'un milliard de fois[2]. Les jeux vidéo sont donc devenus les activités de loisirs les plus sollicitées par les jeunes, ce qui a poussé les chercheurs dans plusieurs disciplines et en particulier dans le domaine de l'éducation à explorer cette tendance et étudier l'exploitation de la puissance cognitive des jeux vidéo pour mieux engager et motiver les élèves dans leurs apprentissages (Gee, 2005 ; Loh, 2009).

La terminologie actuellement utilisée pour les jeux vidéo dont la finalité première est autre que le simple divertissement est le "serious game" (Michael et Chen, 2005). Ce qui a permit aux jeux vidéo de toucher de nouveaux secteurs tel que l'éducation, la santé, la publicité, la formation professionnelle, l'art numérique…etc (Zyda, 2005).

Le terme "serious game" a été proposé en 2002 par l'Américain Benjamin Sawyer pour décrire "America's Army" (Gudmundsen, 2006), jeu diffusé gratuitement sur internet dont l'objective était de valoriser l'armée américaine et le recrutement et la formation des soldats.

Selon Alvarez, les jeux sérieux sont des «Application informatique, dont l'objectif est de combiner à la fois des aspects sérieux tels que, de manière non exhaustive, l'enseignement, l'apprentissage, la communication, ou encore l'information, avec des ressorts ludiques issus du jeu vidéo. Une telle association a donc pour but de s'écarter du simple divertissement.» (Alvarez, 2007).

Dans l'enseignement de la langue, nous pouvons citer les environnements immersifs 3D ou 2D qui permettent aux élèves de développer et d'évaluer leurs compétences culturelles et linguistiques, tels que les environnements "Thélème" (Schmoll, 2011), la "Eonautes"[3], et tactique irakien (Johnson, 2007).

Notre objectif dans cet article est d'explorer les perceptions et les attitudes des futurs enseignants de la langue amazighe envers les apprentissages par les jeux vidéo. Pour ce, nous avions élaboré un questionnaire auquel les vingt enseignants stagiaires de la langue amazighe au Centre Régional des Métiers de l'Education et de la Formation (CRMEF) à Nador, Maroc, ont répondu. Ce questionnaire s'articule autour de 3 axes: leurs expériences avec le jeu vidéo, leurs attitudes envers les jeux vidéo en général, et leurs attitudes envers l'intégration du jeu vidéo en éducation.

---

2 http://expandedramblings.com/index.php/how-many-people-play-angry-birds-infographic/#.U6TAq5QhDnh; consulté le 29/05/2014

3 http://www.eonautes.com/fr/eonautes-serious-game-fle.html;consulté le 30/05/2014

Après des séances d'activités autour de l'utilisation des jeux sérieux en classe menées avec ces futurs enseignants de langue amazighe, un autre questionnaire est distribué, dont les résultats montrent une très bonne conscience sur le potentiel des jeux sérieux, durant les discussions menées avec les étudiants de ce groupe la majorité des enseignants stagiaires envisagent utiliser les jeux dans leurs pratiques enseignantes, néanmoins ils mentionnent que cette innovation pédagogique ne peut être efficace que si le ministère de l'éducation national leur fournit le matériel nécessaire, et que les parents et les acteurs en éducation soient conscient du potentiel des jeux vidéo éducatifs à motiver et engager les apprenants.

2. Perception des enseignants et enseignants stagiaires sur l'utilisation des jeux en classe

Un nombre limité d'articles scientifiques ont étudié les points de vue des enseignants et des futures-enseignants sur l'utilisation des jeux vidéo en classe (Ruggiero, 2013). Schrader, Zheng et Young, par exemple, soulignent que la plupart des enseignants ne voient pas la valeur immédiate de jeux éducatifs en ligne (Schrader *et al*., 2006). En général, les enseignants ne sont pas parfaitement conscients des dimensions pédagogiques de ces jeux, mais plutôt ils les considèrent simplement comme des récompenses. Les résultats de la recherche menée par Can and Cagiltay sur la perception des futures enseignants envers les jeux vidéo éducatifs soulignent que certains participants ont des doutes, notamment sur les questions de gestion de la classe et l'efficacité pédagogique des jeux vidéo commerciaux (Can et Cagiltay, 2006). Dans une autre enquête sur 98 participants (53 enseignants et 45 enseignants stagiaires) réalisée par Pastore et Falvo pour déterminer comment les jeux peuvent efficacement soutenir l'apprentissage, seulement la moitié des participants ont indiqué qu'ils ont, ou avaient, l'intention d'utiliser les jeux dans leur enseignement (Pastore et Falvo, 2010). Selon une enquête récente effectuée par Ruggiero (Ruggiero, 2013), sur une population de 1714 individus (1048 enseignants et 656 enseignants stagiaires), moins de la moitié des enseignants en service utilisent les jeux dans leur enseignement, et plus de 75% des enseignants stagiaires réclament qu'ils veulent apprendre plus sur l'utilisation des jeux vidéo en classe. Une autre étude de Hsu et Chiou (Hsu et Chiou, 2011) montre que la plupart des futurs enseignants qui ont participé à leur enquête, jouent aux jeux vidéo, et ont souligné quelques principaux obstacles dans l'implémentation des jeux en classe, notamment le manque de compétences et de connaissances pour intégrer efficacement les jeux vidéo en classe. Néanmoins, la plupart des recherches ci-dessus agréent que les enseignants ont des perceptions positives concernant l'utilisation des jeux vidéo en classe, et croient que les jeux vidéo éducatifs seront largement utilisés dans l'avenir.

Au Maroc, des études similaires sont quasiment absentes, d'où la présente étude sur la perception des jeux éducatifs chez les enseignants stagiaires de la langue amazighe au CRMEF de Nador. L'identification des perceptions d'enseignants pour les jeux éducatifs est essentielle pour toute intégration future de ces jeux dans le système éducatif marocain. Nous croyons aussi qu'une bonne conscience des enseignants du potentiel des jeux vidéo entraînera certainement leur acceptation comme outils pédagogiques innovants en classe.

3. Expérimentation

L'étude de perception s'est déroulée à travers l'administration de questionnaire en deux temps. En premier lieu, nous avions évalué la perception des enseignants stagiaires avant des séances d'expérimentation avec des jeux éducatifs. Ils ont donc été appelé à remplir un questionnaire sur leur perception envers ce mode d'enseignement (pré-enquête). Ce questionnaire s'articule autour de 3 axes: leurs expériences avec le jeu vidéo, leurs attitudes envers les jeux vidéo en général et leurs attitudes envers l'intégration du jeu vidéo en éducation. Après cette pré-enquête nous avons mené des activités de veille et de sensibilisation sur le "serious game". Nous avions organisé trois séances ou les enseignants stagiaires ont été amenés à jouer à des jeux vidéo à caractère éducatif en ligne. L'objectif était de promouvoir le concept de "serious game" pour l'enseignement des langues et en particulier la langue amazighe. Durant ces séances, il y avait des discussions avec les stagiaires sur les différentes manières et approches pour une meilleure exploitation de ces jeux dans leurs classes. A la fin de ces séances, un deuxième questionnaire a été distribué aux mêmes enseignants stagiaires pour mesurer l'évolution de leur perception envers les jeux vidéo supportant l'apprentissage.

Tous les vingt enseignants stagiaires de la langue amazighe au CRMEF de Nador ont participé à cette expérimentation. Ils ont été aussi introduits à quelques jeux éducatifs d'enseignement de la langue amazighe déjà développés par l'Institut Royal de la Culture AMazighe[4] (IRCAM) (voir figure 1).

4. Résultats

L'analyse des résultats de la pré-enquête a montré que seulement deux étudiants (soit 10%) n'ont jamais joué aux jeux vidéo par manque d'intérêt et/ou du temps. 72,2% des joueurs consacrent au plus 5 heures par semaine aux jeux vidéo. Les jeux de stratégie (avec 65% de joueurs) sont les plus utilisés, puis les jeux de sport (61,2%) et de combat (55 ,6%). Presque 50% des participants attestent que les jeux vidéo n'influencent pas leur rendement scolaire et/ou professionnel.

La mesure de la capacité de leur engagement dans le jeu a montré que 66.6% font de leur mieux pour surmonter les défis et les difficultés dans le jeu vidéo, mais seulement 44,5% ont la confiance de surmonter ces défis et atteindre les objectifs du jeu.

4 http://www.ircam.ma/

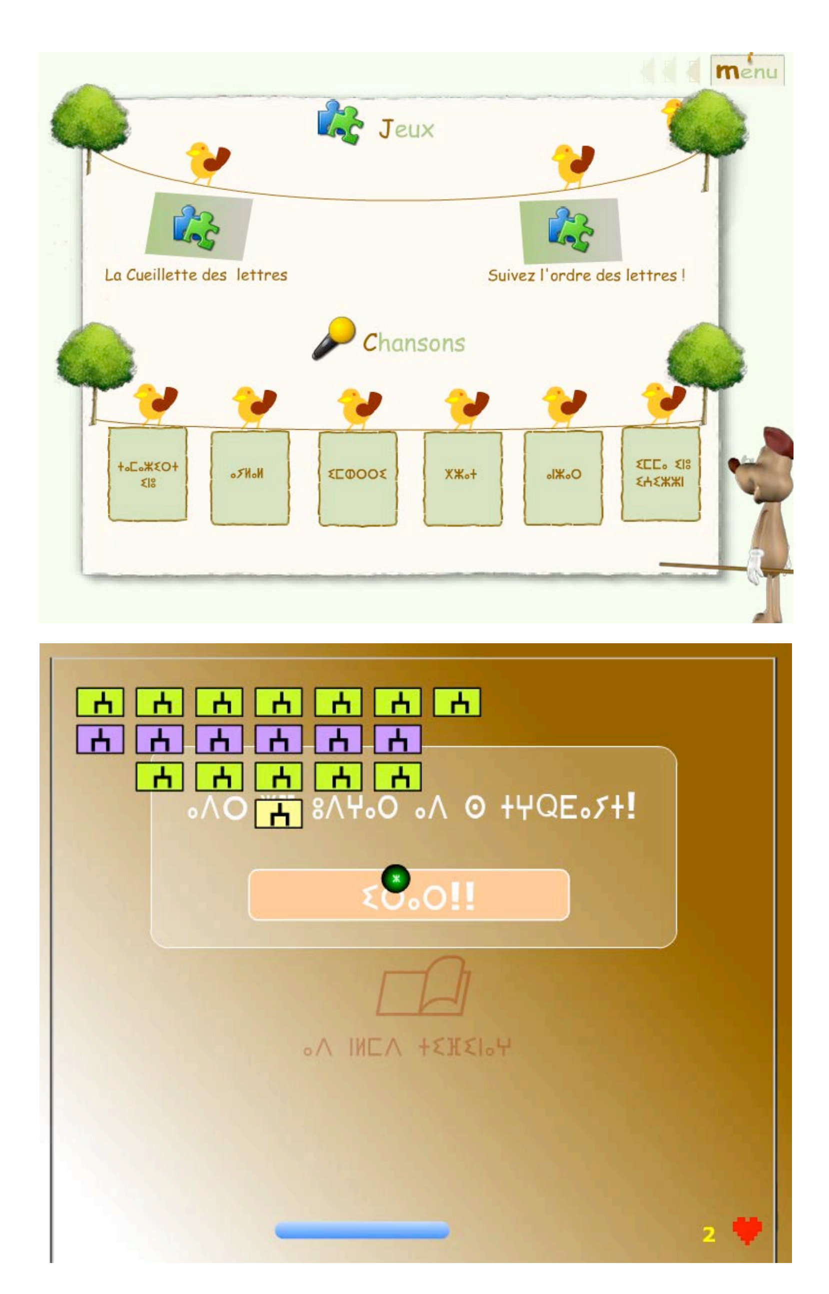


Figure 1 : Exemples de jeux éducatifs développés par l'IRCAM

L'étude de l'évolution de la perception des jeux vidéo comme outil didactique par les enseignants stagiaires de la langue amazighe est résumée dans le tableau suivant :

Question	Avant	Après
Les jeux vidéo éducatifs aident les élèves à développer leur imagination.	85%	100%
Les jeux vidéo éducatifs aident les élèves à développer leurs compétences en communication.	35%	80%
Les jeux vidéo éducatifs peuvent améliorer la capacité de la résolution de problèmes chez les élèves.	45%	85%
Les jeux vidéo éducatifs peuvent aider les élèves à apprendre à coopérer avec les autres.	25%	70%
Les jeux vidéo éducatifs peuvent motiver les élèves pour l'apprentissage	90%	100%
Les jeux vidéo éducatifs peuvent être utilisés comme supports d'apprentissages pour les élèves en classe.	70%	90%
Les jeux vidéo éducatifs peuvent faciliter le travail pédagogique des enseignants en classe.	60%	90%
Je pense utiliser les jeux vidéo éducatifs dans mes pratiques enseignantes	40%	80%

Tableau 1: Étude comparative de l'évolution de la perception envers les jeux vidéo éducatifs avant et après l'expérimentation

Les séances d'introduction des jeux sérieux éducatifs et leur expérimentation a donné l'occasion aux enseignants stagiaires de la langue amazighe d'apprécier l'apport éducatif que peut apporter les jeux éducatifs en classe. La majorité affirment qu'ils ont découvert pour la première fois les ressources numériques sur le site Web de l'IRCAM, ces séances ont aussi permit aux futurs enseignants d'être conscient du potentiel des jeux vidéo pour l'apprentissage, et de penser à innover dans leur pratique enseignante. Les enseignants stagiaires ont soulevé la problématique de la rareté des jeux sérieux pour l'enseignement de la langue amazighe. Même si elles existent, ces ressources ne répondent pas parfaitement aux besoins didactiques nécessaires à enseigner quelques concepts. S'ajoutent aussi, d'après eux, la perception négative des parents et de l'administration aux jeux vidéo comme un support éducatif.

5. Conclusion

Les résultats de cette enquête ont indiqué que la plupart des futurs enseignants, bien qu'ils aient pratiqué le jeu vidéo ("gamer generation"), ignorent quelques aspects d'usage du potentiel de ces ressources dans l'enseignement. Une fois qu'ils ont été initié aux jeux éducatifs, la moyenne de leur perception à considérablement évolué de 56% à 87%. Ceux d'entre eux qui pensent utiliser les jeux éducatifs dans leurs pratiques enseignantes a doublé (de 40% à 80%). Néanmoins, ils pensent que cette innovation pédagogique ne peut être efficace que si elle est appuyée par le ministère de tutelle ou un organisme comme l'IRCAM car ils ne peuvent pas par eux même développer des jeux. Les parents et les acteurs en éducation doivent aussi être sensibilisés sur le potentiel des jeux vidéo éducatifs pour générer la motivation et l'engagement des apprenants.

Les formateurs du CRMEF de Nador ont apprécié cette initiative et soulignent l'innovation que peut générer les jeux vidéo éducatifs dans les pratiques enseignantes de cette nouvelle génération d'enseignants. Nous envisageons par la suite élargir notre expérimentation à un public plus large d'enseignants (stagiaires ou en service) de différentes disciplines.

Références

Alvarez J. (2007). Du jeu vidéo au serious game: approches culturelle, pragmatique et formelle. Thèse de Doctorat, Toulouse 2.

Can G., Cagiltay K. (2006). Turkish prospective teachers' perceptions regarding the use of computer games with educational features. *Educational Technology & Society*, 9(1) :308-321.

Gee J. P. (2005). Good video games and good learning. In *Phi Kappa Phi Forum*, 85(2):33. The Honor Society Of Phi Kappa Phi.

Gudmundsen J. (2006). Movement aims to get serious about games. USA Today. USA. Retrieved from http://www.usatoday.com/tech/gaming/2006-05-19-serious-games_x.htm

Hsu T. Y., Chiou G. F. (2011). Preservice teachers' awareness of digital game-supported learning. *Society for Information Technology & Teacher Education International Conference*, n° 1/2011, pp. 2135-2141.

Johnson W. L. (2007). Serious use of a serious game for language learning. Frontiers in Artificial Intelligence and Applications, Vol. 185, pp. 67-74.

Loh C. S. (2009). Confronting the dark side of video games. *Games: Purpose and Potential in Education*, pp. 185-217. Springer US.

Michael D. R., Chen S. L. (2005). Serious games: Games that educate, train, and inform. Muska & Lipman/Premier-Trade.

Pastore R. S., Falvo D. A. (2010). Video games in the classroom: Pre-and in-service teachers' perceptions of games in the K-12 classroom. *Instructional Technology and Distance Learning*, 7(12):49-61

Ruggiero D. (2013). Video games in the classroom: The teacher point of view. *Games for Learning workshop of the Foundations of Digital Games conference*, Chania, Greece.

Schmoll L. (2011). Usages éducatifs des jeux en ligne: l'exemple de l'apprentissage des langues. *Revue des sciences sociales*, Vol. 45.

Schrader P. G., Zheng D., Young M. (2006). Teachers' perceptions of video games: MMOGs and the future of preservice teacher education. *Innovate: Journal of Online Education*, 2(3):1-10.

Zyda M. (2005). From visual simulation to virtual reality to games. *Computer*, 38(9):25-32.

TICAM'14, p. 49-66

Une approche de construction d'un Wiktionnaire sémantique et multilingue des SHS, en vue de son intégration dans le Web des données ouvertes (Linked Open Data)

Hammou Fadili
Laboratoire CEDRIC du Conservatoire National des Arts et Métiers de Paris
192, rue Saint Martin, 75141, Paris cedex 3, France
Programme Maghreb de la FMSH
190, avenue de France 75013, Paris, France
Hammou.fadili@cnam.fr / fadili@msh-paris.fr

Résumé

Cet article présente une contribution à la construction d'un Wiktionnaire des sciences sociales et humaines (SHS). Cette dernière est basée sur une extension de la plateforme «Mediawiki sémantique» existante afin qu'elle puisse prendre en compte aussi bien les aspects multiculturels et multidisciplinaires des SHS mais aussi pour une représentation de ses entrées selon le standard ISO 1951 relatif aux dictionnaires. Sa construction devrait permettre aux chercheurs des deux rives de la Méditerranée d'échanger et de partager des connaissances dans le domaine des sciences sociales et humaines et cela quelque soit leurs lieux géographiques de travail et/ou de résidence. Sa réalisation a été effectuée suivant des normes et protocoles bien spécifiques, en vue de son intégration dans le Web des données ouvertes (Linked Open Data ou LOD). Après le rappel du contexte du projet et les choix technologiques effectués, nous détaillerons les caractéristiques fonctionnelles et conceptuelles, suivies d'une brève description de l'application développée avant la conclusion.

1. Introduction

Les sciences sociales et humaines jouent un rôle primordial dans la compréhension et l'interprétation du contexte économique, culturel et social dans lequel vivent et agissent les populations. L'évolution de la recherche dans ce domaine passe inévitablement par l'échange et le partage des connaissances entre les chercheurs. A la demande de la DRIC (Direction des relations internationales et de la coopération du ministère français de l'éducation nationale), un séminaire avait réuni en janvier 2002 à la Maison des Sciences de l'Homme (MSH) trente chercheurs maghrébins et français. L'objectif était de réfléchir ensemble sur la situation actuelle de la coopération en sciences sociales et humaines entre la France et les pays du Maghreb, et sur les moyens de la développer. C'est dans ce contexte qu'a été défini et lancé en 2005 le programme FSP-Maghreb.

Le comité de pilotage de ce programma a confié à la Fondation MSH la responsabilité de définir et coordonner, en liaison avec les équipes des 23 groupes de recherche ayant participé au FSP, un projet de construction d'un dictionnaire et d'un contenu multilingue, multidisciplinaire et multiculturel en collaboration, dont le laboratoire CEDRIC du Conservatoire des Arts et Métiers est l'un des principaux partenaires. Une fois réalisé, la plateforme du dictionnaire permettra de favoriser les échanges entre chercheurs maghrébins et leurs partenaires français en mettant en commun un ensemble de savoirs automatiquement exploitables, en Français, Arabe, Amazighe et Anglais, sur les deux cultures et les deux sociétés.

Dire que les contenus doivent être automatiquement exploitables, partageables, etc., cela veut dire qu'ils doivent être mis à disposition dans un contexte de données structurées, sémantiquement annotées et liées. Il s'agit, essentiellement dans notre cas, de définir un schéma de description bien adapté et de faire évoluer les fonctionnalités actuellement offertes par les technologies du Web de données qui est une implémentation pragmatique des évolutions du Web 2.0 (réseaux sociaux) et du Web 3.0 (Web sémantique), afin de prendre en charge les problématiques du domaine des sciences humaines et sociales tels que les aspects multilingues, multidisciplinaires et multiculturels des données. Ceci par l'adaptation et l'utilisation de la technologie « WIKI sémantique » et plus particulièrement de la plateforme Mediawiki sémantique de la fondation Wikimédia ; afin prendre en charge les problématiques liées à la normalisation et à l'intégration du dit dictionnaire dans le Web de données ouvertes (LOD).

L'objectif de cet article est de présenter les technologies utilisées et les travaux réalisés pour la mise en œuvre de l'application du Wiktionnaire sémantique des sciences humaines et sociales (WS-SHS). La Section 1 rappelle le contexte et les motivations du présent travail. La Section 2 décrit les choix technologiques effectués. La Section 3 présente quelques éléments de conception et l'architecture du modèle retenus. La Section 3 montre une vue d'ensemble du système développé. La dernière Section conclut l'article et rappelle quelques perspectives et travaux futurs.

2. Contexte & motivations

Le programme du FSP Maghreb vise à appuyer la recherche en SHS au Maghreb et à développer les échanges entre les chercheurs et institutions de recherche maghrébins et leurs partenaires français dans une logique de réseaux et d'ouverture à l'international, notamment dans un cadre euro-méditerranéen. Il vise également à mettre en commun, par une meilleure maîtrise partagée de l'arabe, de l'amazighe et du français, et par une politique d'incitation à la traduction, un ensemble de savoirs sur les deux cultures et les deux sociétés, et à en assurer la diffusion internationale la plus large possible. Il vise ainsi à développer le débat d'idées entre les deux rives de la Méditerranée, et à conforter une communauté scientifique franco-maghrébine engagée dans un dialogue continu.

Trois objectifs principaux ont été retenus pour y parvenir :

- Mettre en place des réseaux de chercheurs des deux rives de la Méditerranée dans le cadre de programmes de recherche conjointe (réseaux intra-maghrébins et Maghreb-France) ;

- Aider au renforcement et au développement de communautés de jeunes chercheurs en Sciences humaines et sociales au Maghreb, mieux intégrées au sein de la communauté scientifique internationale;
- Développer l'accès à l'information scientifique récente et l'échange de connaissances.

C'est dans ce contexte, que 23 projets thématiques ont été engagés, en plus d'un projet transversal commun à tous les projets, pour la mise en place d'une plateforme de communication, de création et d'échanges entres les chercheurs dans le domaine des SHS. D'un point de vue informatique, cette exigence peut se traduire, par le besoin de mettre en place une « sorte » d'ontologie multilingue, multiculturel et multidisciplinaire du domaine des sciences sociales et humaines (SHS) ; c'est-à-dire d'un espace et d'une base de connaissances du même domaine partagés et reconnus par la communauté des chercheurs Français et Maghrébins.

Cette demande est justifiée par le fait :

- Qu'il n'existe aucun dictionnaire des sciences sociales et humaines franco-maghrébin reflétant l'état de la coopération scientifique et culturelle entre la France et le Maghreb dont les lexiques & concepts pourraient évoluer de manière indépendante ;
- Qu'il y a un grand déséquilibre, d'un point de vue de la disponibilité des ressources numériques, entre le Français et les langues du Maghreb ;
- Les dictionnaires peu nombreux sont incomplets ;
- Les traducteurs existants (également peu nombreux) simplifient la traduction en faisant du « mot-à-mot », i.e. les traductions ne prennent pas en compte tous les aspects liés aux contextes des définitions des concepts ;
- Les corpus potentiels pouvant constituer des sources de données sont également rares ;
- Les travaux sur les nouvelles technologies et la normalisation des données des langues du Maghreb sont encore à leurs débuts ;
- Etc.

Pour y répondre, nous avons fait le choix, d'une solution basée sur un Wiktionnaire sémantique des SHS (WS-SHS), qui est l'objet du présent article.

3. Choix technologiques

La solution proposée pour répondre aux exigences exprimées dans le programme, repose sur des choix technologiques particuliers, que nous proposons de décrire dans ce qui suit.

3.1. Web de données (Web of data) & Données liées ouvertes (Linked Open Data ou LOD)

3.1.1. Le Web de données

Une donnée, non interprétée, est un élément fondamental décrivant un fait brut, devient une information une fois interprétée. Jusqu'à présent, l'interprétation automatique des données est très limitée, car, d'une part, les recherches dans ce domaine sont encore à leur début, et d'autre part, cela nécessite des technologies sophistiquées difficiles à mettre en place. D'où la nécessité de mettre en place des solutions innovantes, simplifiant les technologies existantes comme le Web sémantique, le TAL, etc., pour les rendre réalisables et rendre aussi leurs données exploitables. C'est ce contexte qui a favorisé l'émergence du Web des données (Web of Data).

Le Web de données est bien sûr une notion récente, qui est entrain de se populariser et se généraliser au niveau de la publication et de la diffusion des données « structurées » ou « fortement annotées sémantiquement » sur le Web. Il est constitué d'ensembles liés de données liées ouvertes (Linked Open Data ou LOD), suivant des standards et protocoles normalisés ; facilitant leur interprétation, leur exploitation et leur partage. Il est considéré par certains comme une implémentation du Web sémantique et, par d'autres comme une évolution pragmatique du Web sémantique, très vaste et très difficile à maitriser.

« Le Web des données (Linked Data, en anglais) est une initiative du W3C (Consortium World Wide Web) visant à favoriser la publication de données structurées sur le Web, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un réseau global d'informations. » Source Wikipédia.

3.1.2. Données liées ouvertes

Une donnée ouverte (Open Data) est une donnée mise à disposition dans des formats ouverts et normalisés permettant son exploitation sous des licences ouvertes, libres et gratuites garantissant sa réutilisation dans un contexte international.

Le concept de données liées (Linked Data) est un ensemble de données « structurées » ou «fortement annotées», liées suivant des standards normalisés.

La combinaison des deux notions constitue ce qu'on appelle la notion de données liées ouvertes (Linked Open Data) qui a pour but l'exploitation du Web par les machines et par les êtres humains, faisant abstraction aux spécificités technologiques de toutes les sources de données liées. Cela consiste aussi à répondre à la nécessité de disposer de données légalement "ouvertes", c'est à dire librement (ré-) utilisables par l'utilisateur et pour n'importe quel but.

C'est un moyen puissant pour connecter les utilisateurs à la connaissance et les utilisateurs entre eux **chacun dans sa langue** et dans des environnements hétérogènes et conviviaux.

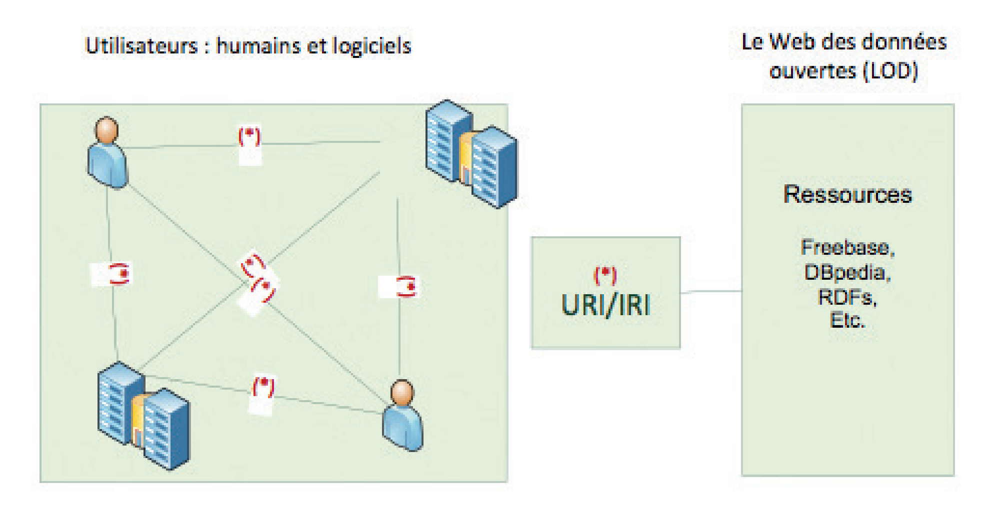


Figure 1 : Les utilisateurs du LOD

3.1.3. Mise à disposition des données dans le Web de données

Le but du Web de données est de permettre une évolution du Web actuel (Web 2.0) où les utilisateurs interagissent avec des sources de données fixes, vers un « Web de données », où les utilisateurs et les machines interagissent potentiellement avec le Web entier pouvant être considéré, à terme, comme une seule base de données universelle, composée des sources de données liées et compatibles. A ce stade de l'évolution de cette nouvelle notion du Web moderne, elle est surtout largement utilisée par les grandes institutions et les gouvernements, pour la mise à disposition de leurs données publiques, dans le cadre de la bonne gouvernance et transparence.

Pour produire et lier les données au Web de données, plusieurs normes, protocoles, standards, etc., sur lesquels doit reposer l'intégration des données, ont été mis en place. Tim Berners-Lee, dans ses travaux, les a résumés à travers les quatre principes suivants :

1. On doit utiliser les URIs comme noms des ressources.
2. On doit utiliser les URIs HTTP accessibles aux utilisateurs.
3. On doit utiliser les normes et standards RDF et SPARQL pour la description et l'interrogation des données.
4. Et puis, on doit créer des liens vers d'autres URIs afin de permettre de découvrir de nouvelles ressources.

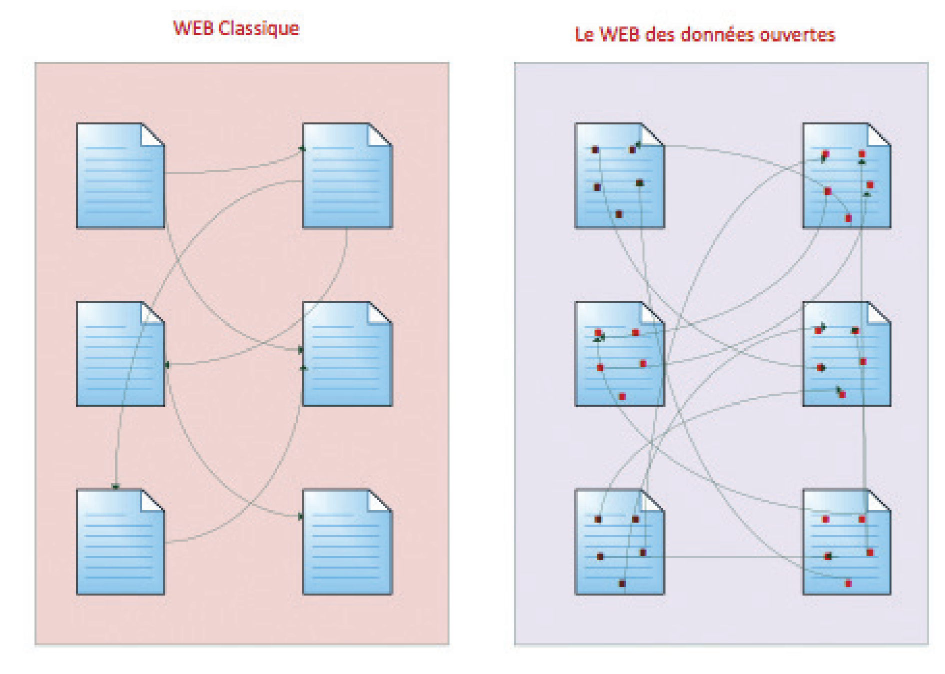


Figure 2 : Evolution du Web vers le LOD

Contrairement au Web classique permettant, seulement la création de liens entre les pages, le Web des données ouvertes permet, en plus, la création de liens entre les données à l'intérieur des pages. Pour mesurer la qualité de la publication des données, Tim Berners-Lee a également défini la classification en nombre d'(*) suivante.

Les conditions associées à chaque ligne sont, à cumuler en plus, de celles des lignes qui précédent.Valeur	
*	Publication des données sous n'importe quel format, mais avec une licence libre (données ouvertes).
* *	Publication des données structurées (RDF, Excel, XML, CSV, etc.).
* * *	Utilisation des formats non propriétaires (exclure Excel et tous les formats propriétaires).
* * * *	Utilisation des standards du W3C pour identifier les ressources (URI, RDF, SPARQ, etc.).
* * * * *	Création de lien vers d'autres ressources (définition du contexte des données).

Figure 3 : Qualité des données dans le LOD

3.2. Wiki sémantique, une implémentation du type « Web sémantique »

3.2.1. Wiki

En 1995, Ward Cunningham a inventé les wikis pour faciliter l'écriture collaborative de contenus.

Un Wiki est en général une application ouverte, universelle et multilingue en ligne. C'est une application de type Système de Gestion de Contenu (CMS) de sites Web permettant à des personnes autorisées, d'éditer et de publier facilement et rapidement des contenus en ligne, de les faire évoluer via des processus de travail collaboratif par mutualisation de compétences, une gestion complète des versions et des historiques des contenus et enfin une gestion des notifications permettant à des personnes intéressées par des thèmes particuliers d'êtres alertées à chaque fois que des créations, modifications ou suppressions de contenus par rapport à leurs thématiques favorites se reproduisent.

La technologie Wiki est basée sur des solutions légères moins coûteuses, accessibles à tous, où l'utilisation ne nécessite pas de formation particulière ; ce qui peut inciter à la création et encourager la production de nouveaux contenus. Ces solutions peuvent être perçues comme outils de communication de groupes, capables de tisser des liens entre communautés en créant des « réseaux sociaux » autour de centres d'intérêts communs. Ces « sociétés virtuelles » peuvent être un levier pour la création de groupes, augmenter la production de la « littérature grise » et ainsi créer une « intelligence collective ». Les Wikis peuvent être aussi considérés comme des outils permettant de sauvegarder la mémoire d'un groupe, d'un projet, d'une institution,… et d'en assurer une meilleure diffusion du fait qu'ils sont très populaires et largement consultés. Les Wiki ont été adaptés et utilisés dans plusieurs domaines : espaces de rédaction, espaces pédagogiques, édition/publication électronique, bibliothèques numériques, Blogs, veille technologique, etc.

3.2.2. Mediawiki

MediaWiki est une plateforme Open source, développé à l'origine pour Wikipédia et aujourd'hui utilisé par plusieurs autres projets de l'association à but non lucratif Wikimedia Foundation ainsi que par beaucoup d'autres institutions privées ou publiques. Wikipédia l'un des projets du Wikimedia Foundation créé en 2001, est devenu le site web le plus visité écrit avec un wiki, développé avec Mediawiki. Néanmoins, le Mediawiki classique et équivalents souffrent de plusieurs limitations, liées surtout à l'interprétation et à l'exploitation de leur contenus par des processus de traitements automatiques. Les contenus de type plein-texte nécessitant des Techniques du Traitement Automatique des Langues (TAL) difficile à mettre en place et dont les résultats de recherche sont encore peu satisfaisants.

3.2.3. Mediawiki sémantique

Pour remédier à ces limitations, et permettre la création et la gestion de contenus compréhensibles et utilisables au même temps par les machines et par les humains, plusieurs solutions basées sur la notion du Web sémantiques ont été mises en place, dont, dans le cas de Wikimedia Foundation, avec l'évolution de Mediawiki vers le Mediawiki sémantique ou

Semantic mediawiki (SMW). Ce dernier a l'avantage de combiner les technologies Wiki et du Web sémantique. Cette extension permet de structurer, rechercher, organiser, naviguer, évaluer et partager des contenus, en les rendant exploitables, aussi bien par les humains que par les machines ; à travers l'utilisation, entre autres, du langage de balises wiki « Wiki-ML » pour les annotations sémantiques, ainsi que beaucoup d'autres extensions annexes.

3.2.4. Semantic forms & pageShema

Semantic forms est une extension du Mediawiki sémantique fournissant des mécanismes de saisie, basés sur des formulaires sémantiques, permettant ainsi la facilité d'édition des données et l'évitement des erreurs de frappes pour la saisie des annotations sémantiques. Cette extension comporte plusieurs utilitaires et outils, comme par exemple, l'introduction de fonctions pour l'auto-complétude des textes, le control du vocabulaire,... La génération de tels formulaires se fait par le biais d'une autre extension appelée « pageSchema » à partir d'un schema XML facile à éditer et paramétrer via un méta-formulaire de définition du schéma.

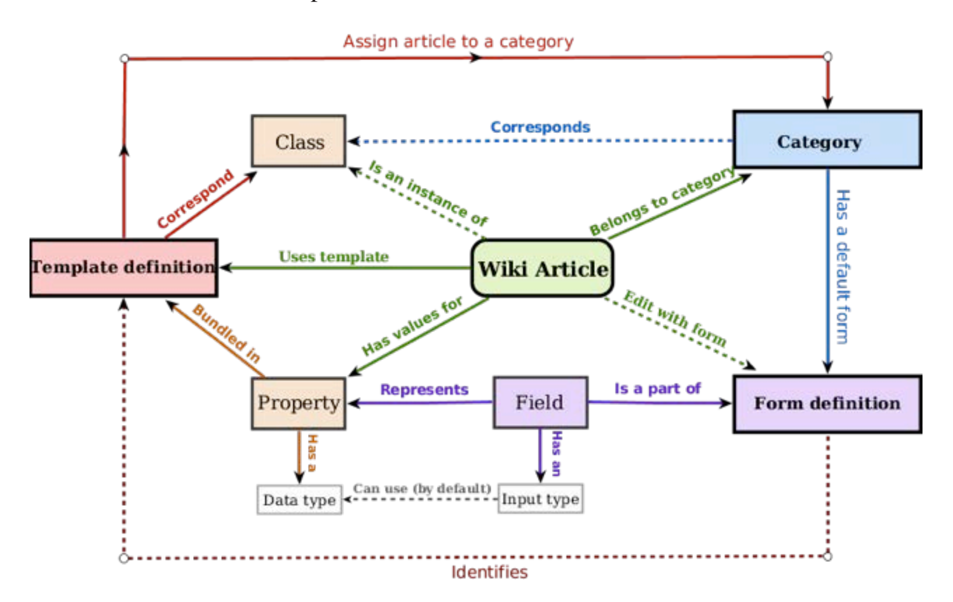


Figure 4 : Meta-modèle « Semantic forms » (Source : EduTechWiki)

C'est une solution idéale pour les besoins du projet, car, cela permet la mise en place d'un Wiktionnaire basé sur un schéma dynamique et extensible « à la demande », en fonction des besoins. C'était une réelle exigence au départ du projet. Etant donné, comme décrit précédemment, que nous devons respecter la norme relative aux dictionnaires ISO 1951, très complexe à mettre en place dans sa totalité, notre choix consistait à prendre en compte un sous-ensemble pertinent du schéma, qu'on peut étendre, « à la demande », en fonction des besoins et disposer ainsi d'une implémentation « POTENTIELLEMENT COMPLETE» de la totalité de la norme ISO des dictionnaires. Cette souplesse a été décisive pour les choix technologiques retenus.

4. Modélisation

4.1. Aspects fonctionnel

L'analyse des besoins nous a permis de déduire un premier modèle fonctionnel de l'application. Le Wiktionnaire des SHS doit :

- Permettre une rédaction distribuée selon les localisations géographiques (France et Maghreb) des différentes équipes participant au programme. C'est l'une des plus importantes fonctionnalités du système ;
- Etre extensible et évolutif, sachant que de nouveaux besoins peuvent apparaître lors de la conception, l'implémentation et l'exploitation du dictionnaire.
- Etre extensible de sorte à englober les langues : anglais, français, arabe et amazighe dans un premier temps puis une extension à toutes les langues du bassin méditerranéen;
- Supporter différents alphabets et graphies des langues retenues.
- Intégrer les différentes disciplines scientifiques du programme FSP, et offrir des passerelles entre celles-ci. Parmi ces disciplines scientifiques, on peut citer l'anthropologie, la sociologie, l'urbanisme, les sciences de l'environnement, etc.
- Etre compatible avec les normes dictionnairiques existantes afin de faciliter les échanges.
- Intégrer les fonctions d'import/export pour la compilation des contenus.
- Permettre l'alimentation de la base dictionnaire dans un format structuré suivant la définition d'une entrée et de sa traduction, et vice versa afin de faciliter l'alimentation, les échanges et la génération des contenus
- Etre doté d'interfaces et de techniques lui assurant une grande ergonomie qui facilitera son utilisation par des chercheurs.
- Prendre en charge l'aspect sémantique des définitions des entrées et de leurs traductions afin de permettre des renvois intelligents d'une langue à une autre, ou encore d'un parler à un autre.
- Etre libre d'accès en consultation, mais réservé en gestion.

4.2. Aspects conceptuels

La conception du Wiktionnaire en SHS repose sur des correspondances entre les éléments de départ dans leurs contextes pour la langue source et les éléments d'arrivée dans leurs contextes pour la/les langue(s) cible(s) selon un sous-ensemble du schéma de la norme ISO1951.

On doit prendre en compte le fait qu'une entrée A_k dans une langue source peut avoir plusieurs sens et donc plusieurs traductions B_1, …, B_j, …B_m dans la langue cible. Cette même entrée Ak peut être définie avec plusieurs éléments A_1, …, A_i, …A_n du schéma du dictionnaire (synonyme, antonyme, étymologie, hyperonyme, hyponyme, etc.) qui peuvent

êtres à leurs tour des entrées dans la même langue source et par conséquent, peuvent avoir plusieurs sens dans cette même langue source et plusieurs traductions dans la langue cible (voir figure 5). Notons que selon le sens de la traduction une langue source peut devenir cible et réciproquement.

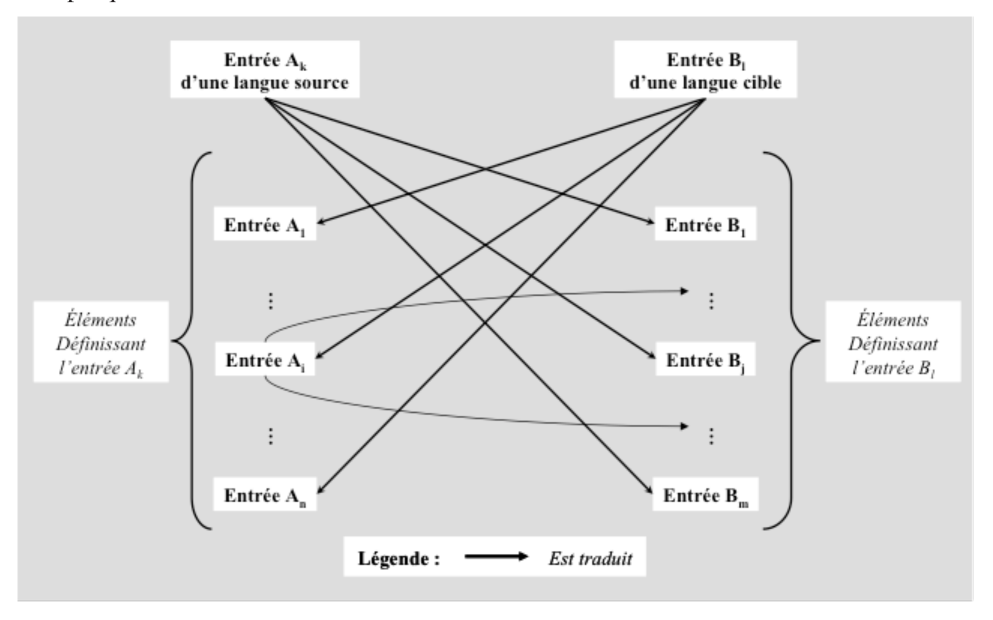


Figure 5 : Extrait du schéma du dictionnaire des SHS

Ce schéma, récapitulant ce qui a été décrit précédemment, montre la complexité des renvois entre la langue source et la langue cible. Elle nécessite la définition d'un schéma spécifique au Wiktionnaire des SHS dans chaque langue ainsi que la mise en place d'un système de gestion des correspondances d'une manière automatique. Pour ce faire, nous pouvons procéder tout d'abord par une première simplification du problème de départ, qui consiste à associer à une entrée source (mot, locution, etc.) un ou plusieurs sens (définitions) qui renvoient à une ou plusieurs entrées cibles ; puis revenir du terme traduit, pris cette fois-ci comme entrée source. Ce procédé doit être pris en charge par la mise en place d'un système de guidage d'aide à la génération et la définition des correspondances entre les entrées, leurs définitions et leurs traductions dans les différentes langues et que l'utilisateur peut compléter à partir des suggestions du système suivant les critères suivants :

- La définition d'une entrée se fait par la description d'une fiche suivant un format structuré déterminé par le schéma du Wiktionnaire sémantique des SHS et composé de: contexte, définitions, relations sémantiques, traductions, indications grammaticales, parlers, etc.
- La signification attribuée à une entrée dépend du contexte de définition. Ce dernier est décrit par un ensemble fini et connu de paramètres contextuels des aspects : temporels, géographiques, disciplinaires, culturels, linguistiques, etc.

- Les relations entre les termes, se fait par le biais de relations sémantiques telles que : la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, l'isonymie, la conversion, etc.
- L'ensemble des éléments servant à décrire une entrée fait partie de la norme ISO 1951 [ISO 1951].

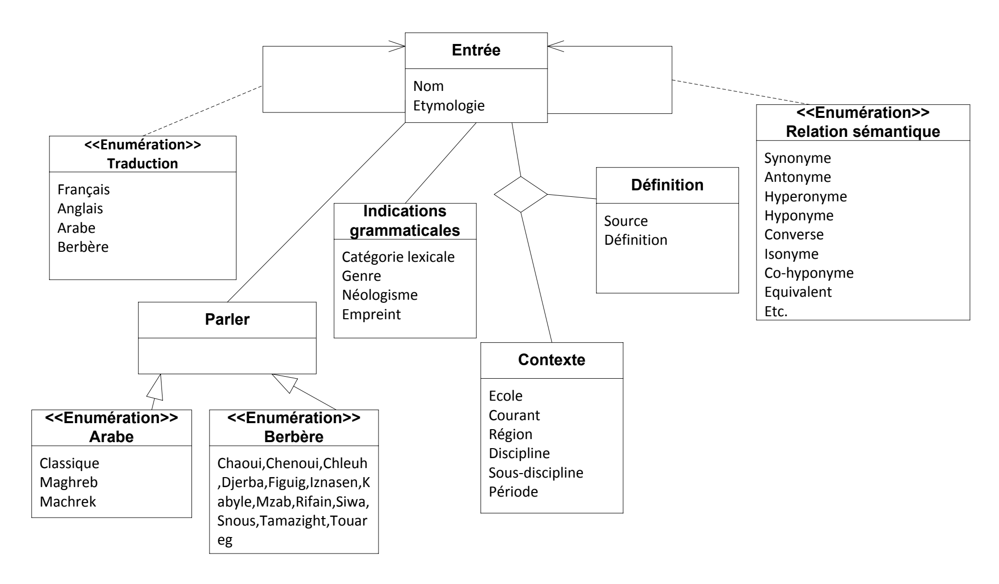


Figure 6 : Modèle du schéma du Wiktionnaire

Pour la mise en place du dictionnaire, nous nous sommes basés sur un sous-ensemble de la norme ISO mentionnée précédemment pour retenir un schéma de base, à partir duquel, on a défini ou traduit des versions pour chacune des langues. Concrètement, nous avons créé deux versions pour chaque formulaire, une en Français et une traduite dans la langue qu'il représente : Français, Arabe, Amazighe ou Anglais. Ceci, permettra, à chacun de renseigner le Wiktionnaire dans sa langue et au non arabophones et aux non amazighophones de le renseigner en utilisant les versions françaises de tous les formulaires.

5. Vue d'ensemble du système développé

Le paragraphe suivant décrit une vue d'ensemble de la plateforme mise en place, à travers 3 sous-paragraphes dédiés respectivement, aux fonctionnalités principales offertes par le système, à nos contributions pour sa mise en place et à un aperçu de l'application à travers quelques copies d'écrans.

5.1. Principales fonctionnalités

L'analyse des besoins nous a permis de déduire un certain ensemble de fonctionnalités importantes que nous devons retenir et implémenter dans la plateforme. A rappeler que

l’application été conçue et développée dans le but d’être compatible et intégrable dans le web de données et dont les données doivent être gérées dans un esprit de partage et coopération. Ci-après une liste non exhaustive des principales fonctionnalités et services implémentés et mis à disposition par la plateforme.

Edition et gestion :

- Gestion ergonomique, collaborative et orientée communautés ;
- Sécurisation des accès et des contenus ;
- Saisie et annotation des données via des formulaires ;
- Gestion automatique de liens ;
- Génération automatique de listes ;
- Complétion automatique de la saisie.

Structuration :

- Catégorisation et hiérarchisation des données ;
- Instanciation des catégories et création des articles ;
- Annotations et création des propriétés et relations des articles ;
- Génération automatique « à la demande » des versions multilingues du schéma ;
- Extension automatique « à la demande » du schéma à d’autres éléments de la norme.

Interrogation/navigation :

- Interrogation sémantique des contenus ;
- Navigation par ontologie ;
- Recherche à facettes ;
- Export/import RDF ;
- Exposition d’un point de terminaison SAPRQL ;
- Intégration dans le Web de données (triplets RDF) ;
- Navigation riche entre les langues.

Diffusion et enrichissement automatique du dictionnaire :

- Meilleure diffusion : l’intégration du dictionnaire dans le Web de données, permet bien sûr une meilleure diffusion ; les sources de données qui lui sont liées sont dans la plupart des cas très connues et populaires, comme par exemple : le dbpedia.
- Enrichissement automatique : nous avons créé pour cela des requêtes paramétrées SPARQL permettant, suivant le contexte, de générer des contenus bien spécifiques, soit pour créer de nouvelles pages, soit pour enrichir d’une manière automatique la définition d’un concept SHS du dictionnaire. Le Wiktionnaire peut se servir

potentiellement de la totalité du Web de données compatibles pour enrichir ses contenus.

5.2. Contribution et améliorations

L'implémentation de l'application a été réalisée à partir de la version de base de la plateforme Mediawiki sémantique et de ses extensions. Nous avons procédé à des développements, améliorations et adaptations des différents outils, que l'on peut résumer comme suit :

- Développement et adaptation du module associé au méta-formulaire basé sur «pageSchema », pour la génération automatique des formulaires du module « semantic forms », et la prise en charge de toutes les langues dont l'arabe. Les versions arabes de ces modules n'existent pas.
- Développement et adaptation du module « semantic forms » pour la génération de modèles d'affichage multilingues dont le cas de l'arabe.
- Personnalisation des feuilles de styles en fonction de la langue affichée.
- Définition d'un modèle pertinent du schéma du dictionnaire :
  - Basé sur des catégories, propriétés, relations,...
  - Définition des versions spécifiques du schéma de base, et ses versions légèrement différentes pour l'Arabe et l'Amazighe
- Définition et génération de tous les formulaires.

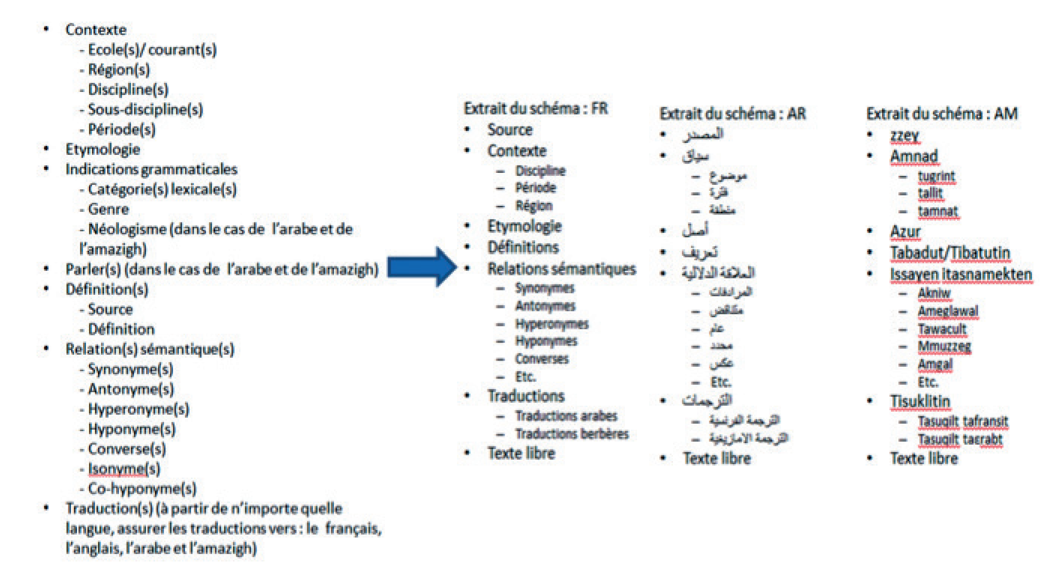


Figure 7 : Adaptation d'un extrait du schéma du dictionnaire

5.3. L'application

L'application est issue d'une adaptation de la technologie du MédiaWiki sémantique et de ses extensions pour prendre en charge la totalité des besoins de gestion du sous-ensemble du schéma du Wiktionnaire retenu dans les différentes langues. Ce dernier doit être facilement et rapidement extensible « à la demande », et doit couvrir, potentiellement, la totalité de la norme ISO1951. Nous avons pris en compte les spécificités de chaque langue, en particulier de l'arabe et son intégration dans le modèle de saisie orienté formulaire, comme par exemple l'écriture de la droite vers la gauche. La solution, a l'avantage de nous offrir la possibilité de réaliser, dans un premier temps un Wiktionnaire anglais-français-arabe-amazighe, facilement extensible à d'autres langues pratiqués dans le bassin méditerranéen, contrainte majeure d'implémentation spécifiée dans les besoins initiaux du projet.

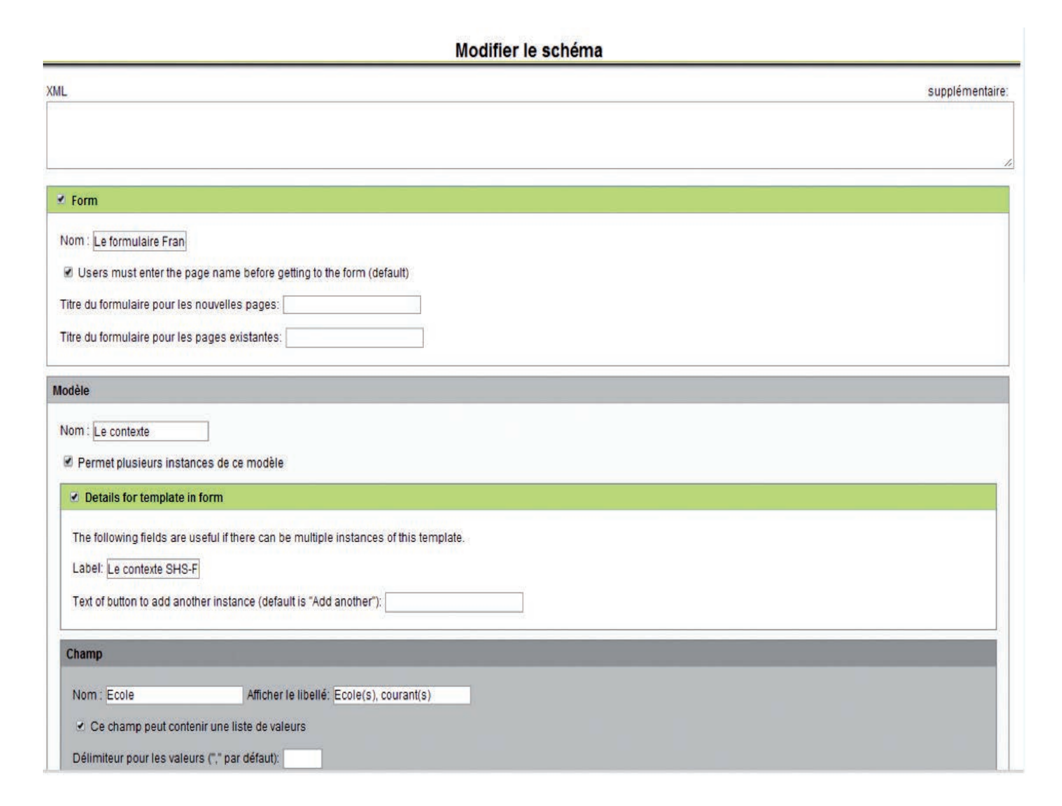
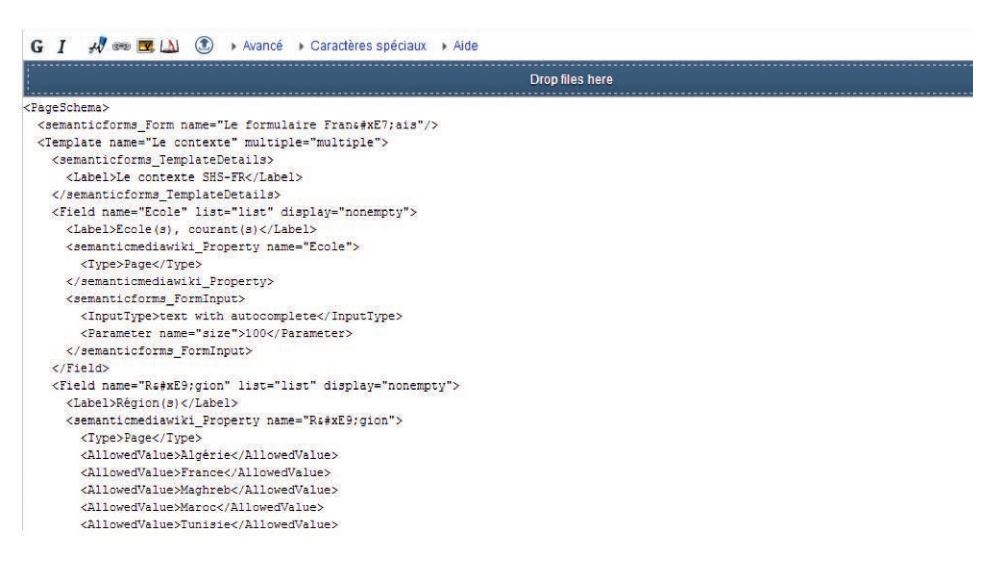


Figure 8 : Formulaire de gestion des schémas

Le méta-formulaire de création du schéma du Wiktionnaire permettant la génération automatique des formulaires de saisie ou d'annotation sémantique.


G I Avancé Caractères spéciaux Aide
Drop files here

```
<PageSchema>
  <semanticforms_Form name="Le formulaire Fran&#xE7;ais"/>
  <Template name="Le contexte" multiple="multiple">
    <semanticforms_TemplateDetails>
      <Label>Le contexte SHS-FR</Label>
    </semanticforms_TemplateDetails>
    <Field name="Ecole" list="list" display="nonempty">
      <Label>Ecole(s), courant(s)</Label>
      <semanticmediawiki_Property name="Ecole">
        <Type>Page</Type>
      </semanticmediawiki_Property>
      <semanticforms_FormInput>
        <InputType>text with autocomplete</InputType>
        <Parameter name="size">100</Parameter>
      </semanticforms_FormInput>
    </Field>
    <Field name="R&#xE9;gion" list="list" display="nonempty">
      <Label>Région(s)</Label>
      <semanticmediawiki_Property name="R&#xE9;gion">
        <Type>Page</Type>
        <AllowedValue>Algérie</AllowedValue>
        <AllowedValue>France</AllowedValue>
        <AllowedValue>Maghreb</AllowedValue>
        <AllowedValue>Maroc</AllowedValue>
        <AllowedValue>Tunisie</AllowedValue>
```

Figure 9 : Version XML générée du schéma

Version XML du schéma généré par le méta-formulaire de gestion de schémas précédent.

Modifier Le formulaire Français : Entrepreneur(1)

La(es) source(s) SHS-FR

La(es) source(s): Auteur1

Ajouter une autre

Le contexte SHS-FR

La(es) région(s): ☑ France ☑ Maroc ☐ Algérie ☐ Tunisie ☐ Maghreb
La(es) discipline(s): ☑ Histoire ☑ sociologie ☐ droit
La(es) période(s): ☑ 17eme ☑ 18eme ☑ 19eme ☐ 20eme ☐ 21eme ☐ 15ème ☐ 16ème

Ajouter une autre

Ethymologie

L'éthymologie: Ethy1, Ethy2,

Ajouter une autre

Définition FR

Définition 1:
On parle également de l'économie lato sensu comme de la situation économique d'un pays ou d'une zone, c'est-à-dire de sa position conjoncturelle (par rapport aux cycles économiques) ou structurelle. Dans ce sens, l'économie est donc un quasi-synonyme à la fois de système et de régime. Enfin, de manière générale, en français, on parle d'économie comme synonyme de réduction de dépense ou d'épargne. L'économie peut en effet être le résultat d'une organisation interne plus efficiente : on parle alors d'économie interne. La baisse du coût moyen due à l'augmentation de la dimension de l'entreprise constitue une économie d'échelle ou économie

Définition 2:
On parle également de l'économie lato sensu comme de la situation économique d'un pays ou d'une zone, c'est-à-dire de sa position conjoncturelle (par rapport aux cycles économiques) ou structurelle. Dans ce sens, l'économie est donc un quasi-synonyme à la fois de système et de régime. Enfin, de manière générale, en français, on parle d'économie comme synonyme de réduction de dépense ou d'épargne. L'économie peut en effet être le résultat d'une organisation interne plus efficiente : on parle alors d'économie

Figure 10: Formulaire d'édition.

Formulaire d'édition implémentant le schéma du Wiktionnaire retenu à un moment donné pour une langue donnée.

Entrepreneur(1)

La(es) source(s)

La(es) source(s)	Auteur1

Le contexte

La(es) région(s)	France, Maroc
La(es) discipline(s)	Histoire, sociologie
La(es) période(s)	17eme, 18eme, 19eme

La(es) Ethymologie(s)

L'éthymologie	Ethy1, Ethy2

La(es) Définition(s) FR

Définition 1	On parle également de l'économie lato sensu comme de la situation économique d'un pays ou d'une zone, c'est-à-dire de sa position conjoncturelle (par rapport aux cycles économiques) ou structurelle. Dans ce sens, l'économie est donc un quasi-synonyme à la fois de système et de régime. Enfin, de manière générale, en français, on parle d'économie comme synonyme de réduction de dépense ou d'épargne. L'économie peut en effet être le résultat d'une organisation interne plus efficiente : on parle alors d'économie interne. La baisse du coût moyen due à l'augmentation de la dimension de l'entreprise constitue une économie d'échelle ou économie de dimension. L'économie peut résulter d'un phénomène extérieur au pouvoir de décision de l'agent : on parle alors d'économie externe ou externalités qui peuvent être soit positives, si elles apportent un plus aux agents économiques, ou négatives, dans le cas contraire. On parle également de l'économie lato sensu comme de la situation économique d'un pays ou d'une zone, c'est-à-dire de sa position conjoncturelle (par rapport aux cycles économiques) ou structurelle. Dans ce sens, l'économie est donc un quasi-synonyme à la fois de système et de régime. Enfin, de manière générale, en français, on parle d'économie comme synonyme de réduction de dépense ou d'épargne. L'économie peut en effet être le résultat d'une organisation interne plus efficiente : on parle alors d'économie interne. La baisse du coût moyen due à l'augmentation de la dimension de l'entreprise constitue une économie d'échelle ou économie de dimension. L'économie peut résulter d'un phénomène extérieur au pouvoir de décision de l'agent : on parle alors d'économie externe ou externalités qui peuvent être soit positives, si elles apportent un plus aux agents économiques, ou négatives, dans le cas contraire.
Définition 2	On parle également de l'économie lato sensu comme de la situation économique d'un pays ou d'une zone, c'est-à-dire de sa position conjoncturelle (par rapport aux cycles économiques) ou structurelle. Dans ce sens, l'économie est donc un quasi-synonyme à la fois de système et de régime. Enfin, de manière générale, en français, on parle d'économie comme synonyme de réduction de dépense ou d'épargne. L'économie peut en effet être le résultat d'une organisation interne plus efficiente : on parle alors d'économie interne. La baisse du coût moyen due à l'augmentation de la dimension de l'entreprise constitue une économie d'échelle ou économie de dimension. L'économie peut résulter d'un phénomène extérieur au pouvoir de décision de l'agent : on parle alors d'économie externe ou externalités qui peuvent être soit positives, si elles apportent un plus aux agents économiques, ou négatives, dans le cas contraire.

Figure 11 : Interface de navigation

Interface de lecture des entrées associées à des articles annotés suivant les modèles de description associés à chaque langue.

Point de terminaison SPARQL

Enter SPARQL query:

```
@PREFIX w : <Spécial:URIResolver/> .

SELECT *
WHERE { ?s ?p ?o }
LIMIT 25
```

Query by Equivalent URIs: Output Equivalent URIs: Output Equivalent

Output format: HTML

(RDF/XML requires creating triples using CONSTRUCT)

Submit

Result:

s	p	o
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#DatatypeProperty
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://www.w3.org/2000/01/rdf-schema#label	Discipline(s)
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://semantic-mediawiki.org/swivt/1.0#page	http://france-maghreb.msh-paris.fr/wiktionnaire/index.php/Attribut:Dis
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://www.w3.org/2000/01/rdf-schema#isDefinedBy	http://france-maghreb.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://semantic-mediawiki.org/swivt/1.0#wikiNamespace	102
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://semantic-mediawiki.org/swivt/1.0#wikiPageCreationDate	2013-07-23T10:17:37Z
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADate_de_cr-C3-A9ation-23aux	2456496.9289
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://semantic-mediawiki.org/swivt/1.0#wikiPageLastEditor	http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIR
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://semantic-mediawiki.org/swivt/1.0#wikiPageModificationDate	2013-09-17T16:08:38Z
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADate_de_modification-23aux	2456553.17266
http://www.msh-paris.fr/wiktionnaire/index.php/Sp%C3%A9cial:URIResolver/Attribut-3ADiscipline-28s-29	http://semantic-mediawiki.org/swivt/1.0#specialProperty_PVAL	Histoire

Figure 12 : Point de terminaison et interrogation SPARQL

Interface de terminaison et d'interrogation SPARQL.

6. Conclusion et travaux futurs

Nous avons présenté dans cet article une approche de construction d'un dictionnaire multilingue, multidisciplinaire et multiculturel franco-maghrébin pour les sciences sociales et humaines. Ce projet a été initié par la Fondation Maison des Sciences de l'Homme.

L'architecture mise en place repose sur une conception générique représentant le choix du schéma et des langues retenus à un moment donné, mais facilement extensible à d'autres schémas et généralisable à d'autres langues. La plateforme développée est une extension de la plateforme Mediawiki et plus particulièrement de la plateforme Mediawiki sémantique, garantissant une utilisation ergonomique : saisie, annotations, consultation, navigation et recherche sémantiques. Elle est en phase de production après plusieurs mois de développements-tests-améliorations et une validation par un comité d'experts/chercheurs du domaine.

Comme perspective de recherche à ce travail, on prévoit, l'intégration du Wiktionnaire dans le Web de données, puis l'utilisation des données, structurées, sémantiquement annotées et liées (intégrées dans le Web de données) pour enrichir un travail de recherche parallèle, concernant une nouvelle approche de détection automatique et contextuelle du sens dans le texte : Basé sur les relations lexico-sémantiques et la notion de contexte.

Bibliographie

http://www.w3.org/DesignIssues/LinkedData.html
http://linkeddata.org/
http://linguistics.okfn.org/2011/05/20/the-open-linguistics-working-group/
http://ldl2012.lod2.eu/
http://inkdroid.org/lod-graph/
http://www.mediawiki.org/wiki/Special:ExtensionDistributor
https://semantic-mediawiki.org/

Towards an Amazigh Geo-cultural Information system

Ali Rachidi Mustapha Eddahibi
Equipe de recherché TISED, Ecole Nationale de Commerce et de Gestion, Agadir, Maroc
{a.rachidi,m.eddahibi}@uiz.ac.ma

Abstract

In this paper, we will give a short overview of the work that has been started for the conception and the implementation of a new Amazigh geo-cultural information system "AGCIS". The aim of this work is to implement a geographical information system about Moroccan Amazigh culture in its various aspects, to preserve the cultural heritage, and to promote Amazigh language and culture. At this initial stage, we wish to find support amongst organisms that holds the official information about Amazigh culture.

1. Introduction

A Geo-cultural information system is a computer platform that allows creating, displaying, searching, processing and storing cultural resources according to each geographical area. AGCIS associates spatial data with cultural resources. It will be used to produce cultural documents, maps and every kind of information useful in touristic field, culture promotion, and discovery:

- Data storage and management in a data base;
- Data processing and data querying to respond to every request;
- Formatting, visualization, printing and exporting data, depending on the need;
- Representing the real Amazigh world and without abstraction;
- Anticipating Amazigh cultural new development in Morocco.

Therefore, the architecture of AGCIS system can be presented as follows in Figure1.

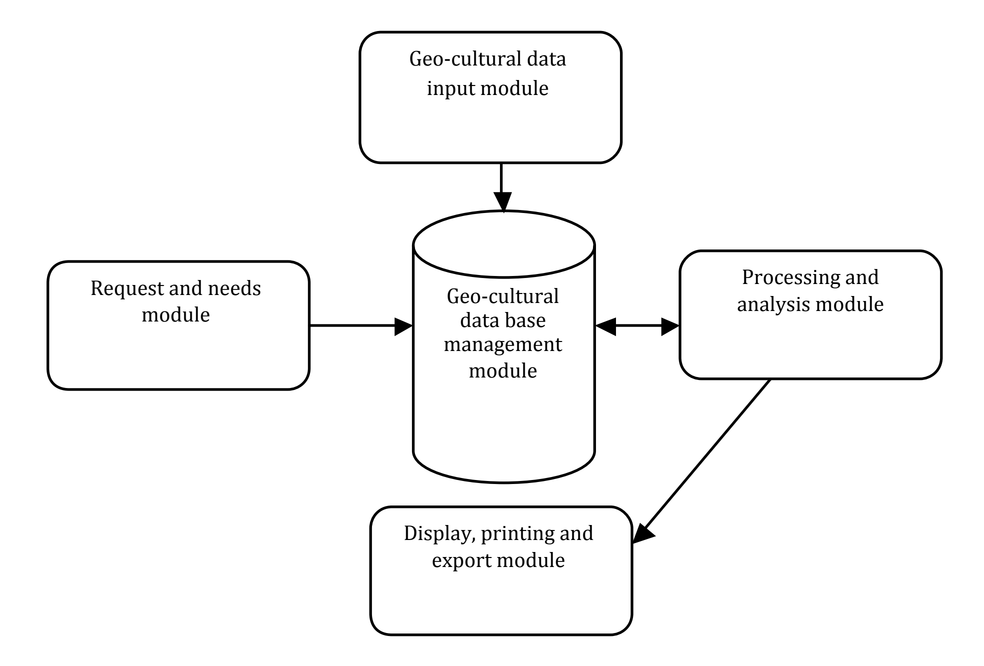


Figure 1: Amazigh geo-cultural information system architecture

Indeed, two categories of data can be entered in the AGCIS's database:

- Spatial data using geographical maps for different Morocco's regions. These maps are based on satellite images available in the web or from governmental institutions as the High Commissioner for the Plan, urban authorities' cadaster service, etc.
- Amazigh cultural data describing different aspects of Amazigh culture for each region. Cultural data would be strongly bound to their related geographical data. All data are available in the culture ministry, the ministry of tourism, IRCAM, etc.

2. Amazigh Culture and its aspects

Moroccan Amazigh people have their own culture, their particular language, their own history, their specific traditions and customs that characterize them and distinguish them from other social groups. Through ongoing contact with nearby communities and through the process of acculturation, they have been directly involved in the advancement of the Moroccan human civilization. Indeed, most social scientists today view culture as consisting primarily of the symbolic, ideational, and intangible aspects of human societies. The essence of a culture is not its artifacts, tools, or other tangible cultural elements but how the members of the group interpret, use, and perceive them. It is the values, symbols, interpretations, and perspectives that distinguish one people from another in modernized societies; it is not material objects and other tangible aspects of human societies. People within a culture usually interpret the meaning of symbols, artifacts, and behaviors in the same or in similar ways. Culture includes every aspect of life: know-how, technical knowledge, customs of food and dress, religion,

mentality, values, language, symbols, socio-political and economic behavior, indigenous methods of taking decisions and exercising power, methods of production and economic relations, and so on.

This culture can be decomposed by audience (individual or group) and / or by type (subjective or objective). Table 1 presents the decomposition adopted in our work.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Topic</td></tr>
<tr><td>Subjective</td><td>Objective</td></tr>
<tr><td rowspan="2">Audience</td><td>Individual</td><td>Intention
• Values
• Attitudes
• Commitment
• Responsability</td><td>Behaviour
• Plans
• Actions
• Decision
• Performance</td></tr>
<tr><td>Group</td><td>Culture
• Shared Values
• Ethics
• Morals</td><td>Systems
• Organizational structures
• Work Process
• Polices
• Procedures
• Contracts</td></tr>
</table>

Table 1: Culture presentation model

A database of culture is presented in the section of development.

3. Development and implementation

To realize our system, it is necessary, first of all, to develop the cultural database. Initially, we adopted entities – relationships model. The result model is developed with MS Access. It is expected, in the course of our work, to migrate this database to MYSQL system. Figure 2 presents the model obtained.

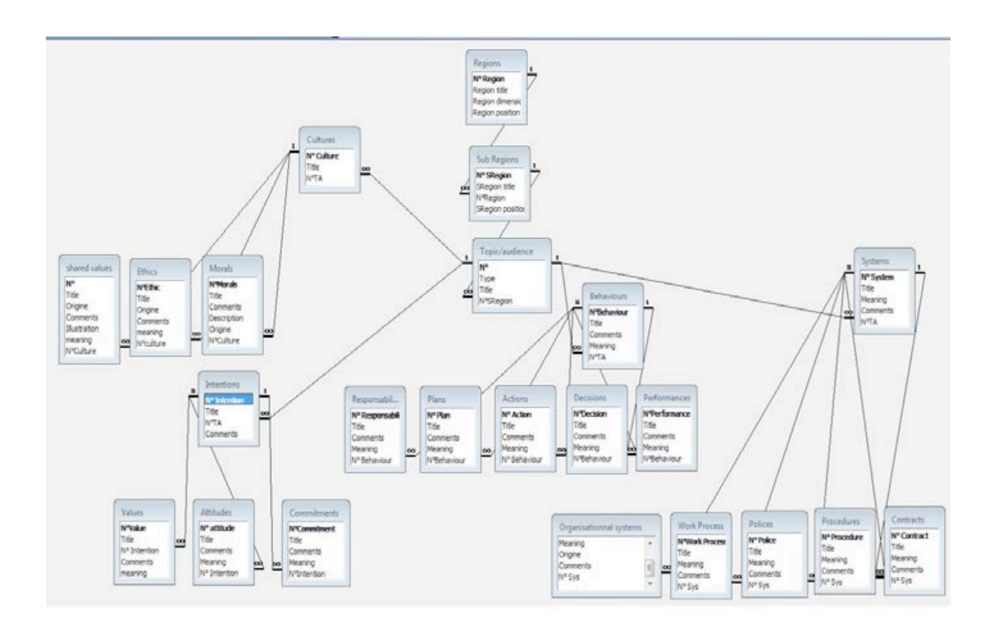


Figure 2: Database of Amazigh geo-cultural information system

For the development and the implementation of the Amazigh geo-cultural information system, we are using the Open Source Geographic Information System QGIS that operates within open Spatial Data Infrastructure. It provides to share and edit geospatial data.

Up to now, we are using only one format for the storage of geospatial data which are shapefiles. A shapefile is a digital vector storage format for storing geometric location and associated attribute information. Shapefiles describe spatially vector features like points, lines, and polygons. A table of records is used to store properties for each shape in the shapefile. Associating shapes with data attributes create many representations about geographic data. To publish Shapefiles with QGIS, we first create a workspace for this shapefile. The workspace is a container used to group similar layers together. Then, we create data store that tells QGIS how to connect to the database. Finally, we create a Layer by defining the Data publishing parameters. Layer's bounding boxes are then computed from native data. Figure 3 presents a shapefile developed in QGIS.

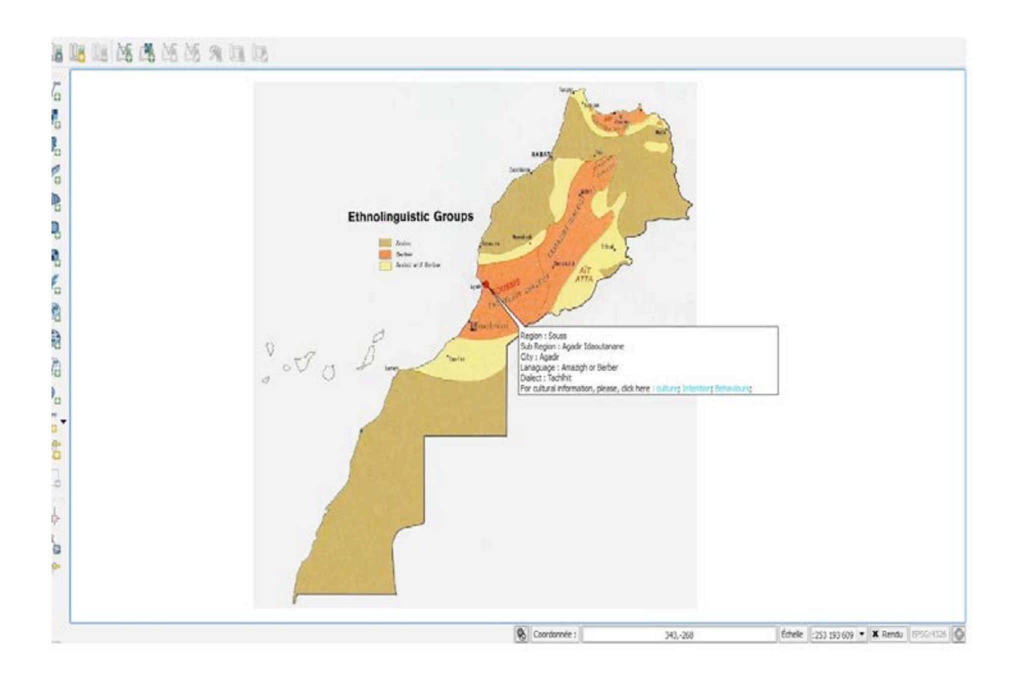


Figure 3: Shapefile of AGCIS

4. Conclusion

In this paper, we have presented an Amzigh geo-cultural information system that we have initiated in the research team TISED. This system would ensure Amazigh cultural information storage for each region and location in Morocco. Therefore, it would contribute to the conservation of Amazigh culture, and to promote it. Furthermore, this tool will be free, open source and available on the web to be widely exploited by the world wide community for noble purposes and for major interest vocation in Morocco such as tourism, etc.

Bibliography

Chafik M. (1989). Lamha âan thalatha wa thalathine qarnan min tarikh alamazighyine (Aperçu sur trente-trois siècles de l'histoire des Amazighs).

Kappos A., Rivard S. (2008). A three-perspective model of culture, information systems, and their development and use. *MIS Quarterly*, 32(3):601-634.

Stephen J. (2011). Organizational culture and information systems adoption: A three-perspective approach. *Information and Organization*, 21(2):57-83.

TICAM'14, p. 73-80

Mesure de l'équivalence lexicographique entre les variantes de la langue Tamazight

Ali El Houssaini [1] **Benaceur Outtaj** [2]
[1] ETSE, ENSIAS, Université Mohammed V, Rabat
ali.elhoussaini@um5s.net.ma
[2] FSJES, Souissi, Université Mohammed V, Rabat
b.outtaj@um5s.net.ma

Résumé

Dans le but de contribuer à la normalisation de la langue Tamazight, nous étudions dans cet article, les similarités linguistiques entre les différentes variantes de la langue Tamazight (Tamazight, Tachelhit, et Tarifit) en employant l'une des distances le plus réputées dans le domaine de traitement automatique des langues : la distance de Levenshtein. Les premiers résultats issus de l'application de cette distance sur un corpus modeste, construit à base de mots équivalents (du point de vue lexicographique), montrent que la similarité est très forte. Ce qui nous conduit à confirmer les hypothèses formulées en matière d'équivalence linguistique ou phonétique sur certains caractères ou phonèmes.

1. Introduction

L'intégration de la langue Tamazight dans le domaine des nouvelles technologies de l'information et de la communication (NTIC) suscite ces derniers temps un intérêt considérable de la part de tous les chercheurs qui défendent la promotion de cette langue. Parmi les aspects les plus intéressants, on distingue le chantier de normalisation entrepris par l'IRCAM (Ameur et Boumalk, 2004). Cependant, les travaux effectués jusqu'à présent dans ce domaine se concentrent plus sur l'aspect linguistique et n'intègrent pas les NTIC pour aider, voire accélérer le processus de normalisation de la langue Tamazight.

Dans ce cadre, et dans le but de contribuer au processus de normalisation, nous présentons dans ce papier une étude que nous avons réalisée pour mesurer d'une manière automatique (i.e. en utilisant l'outil informatique) l'équivalence lexicographique entre les trois « différentes » variantes de la langue Tamazight à savoir : Le Tachelhit parlée au sud du Maroc, le Tamazight parlée au moyen Atlas et le Tarifit utilisée au Nord du Maroc.

Dans le domaine du traitement automatique des langues, on utilise souvent la notion de «distance» pour mesurer les similarités ou les différences entre les mots d'une langue. Et comme dans le cas de la langue Tamazight, le même mot (du point de vue lexicographique) se prononce différemment dans les trois variantes (et donc s'écrit différemment), nous avons pensé à mesurer ces différences en vue d'établir ou de confirmer des assertions d'équivalences linguistiques ou phonétiques entre certains caractères ou phonèmes de notre langue.

Le présent article est organisé comme suit : La Section 2 rappelle la méthode utilisée dans la mesure de ces équivalences pour justifier son choix et usage. Dans la Section 3, nous appliquons la distance de Levenshtein (Levenshtein, 1966).pour quantifier ces mesures. Alors que la dans le Section 4, nous formulons quelques conclusions à propos des équivalences. La Section 5 dresse une conclusion de notre étude et les perspectives envisagées surtout en matière d'extension de notre travail.

2. Méthode utilisée

Dans le traitement automatique des langues naturelles, des notions mathématiques et algorithmiques ont été introduites pour mesurer la « ressemblance », la « similarité » ou «l'équivalence» entre deux ou plusieurs mots pour essayer de répondre plus particulièrement aux besoins de correction orthographique des erreurs de saisie. Ces méthodes ont contribué amplement dans l'amélioration du niveau de maîtrise des langues tant au niveau de l'écriture qu'au niveau de l'usage oral (Kukich, 1992). La notion de distance, au sens mathématique, a été revue et améliorée pour qu'elle couvre les langues naturelles et intègre l'utilisation des NTIC. Parmi ces techniques, la distance de Levenshtein est l'une des méthodes la plus utilisée dans le domaine du traitement automatique des langues naturelles (TAL) depuis les années 60 et a fait l'objet dernièrement d'une proposition pour la correction automatique des erreurs orthographiques par des collègues chercheurs de notre équipe de recherche (Gueddah *et al*., 2012).

En mathématique et/ou en informatique, une mesure permet de calculer la similarité ou la dissimilarité entre deux chaînes de caractères. Dans le domaine du TAL, ces deux chaînes de caractères ne sont que deux mots d'une langue ou dans notre cas, deux mots de deux variantes linguistiques d'une grande langue mère (Tamazight). La mesure de similarité consiste donc à comparer les mots correspondants. Dans certains cas et suite aux améliorations qui peuvent être apportées sur ces mesures, deux mots qui représentent deux chaînes de caractères différentes peuvent être considérées comme similaires. Par exemple, les chaînes de caractères "Sam" et "Samuel" peuvent être considérées comme similaires (Elsa, 2013). Une telle mesure sur les chaînes de caractères fournit une valeur obtenue algorithmiquement.

La distance de Levenshtein est couramment utilisée pour calculer de telles mesures de similarité appelées encore distance de l'étude d'édition entre deux chaînes.

2.1. Définition :

On appelle la distance de Levenshtein entre deux mots S et T, le coût minimal pour aller de S à T en effectuant les opérations élémentaires suivantes:

- Substitution d'un caractère de S par un caractère de T.
- Ajout dans S d'un caractère de T.
- Suppression d'un caractère de T.

On défini un coût de substitution pour tout les couples de symboles, ainsi qu'un coût d'insertion et de suppression pour chacun de ces symboles. Dans le cas le plus simple, le coût de substitution vaut « 1 » lorsque les symboles sont identiques et « 0 » s'ils sont différents, les coûts d'insertion et de suppression sont également égaux à « 1 ».

2.2. Fonctionnement :

Soit « **s** » la longueur de la chaîne **S** et « **t** » la longueur de la chaîne **T**. Le pseudo-code de l'algorithme qui nous permet de calculer cette distance est présenté ci-dessous en deux étapes, avec des techniques de programmation dynamique. On peut le calculer en temps quadratique et linéaire en espace (Wechsler *et al.*, 1998).

Première étape :

Construire une matrice M de s + 1 lignes et de t + 1 colonnes.

Initialiser la première ligne à [0, 1, 2, ..., s].

Initialiser la première colonne à [0, 1, 2, ..., t].

Deuxième étape :

Pour i = 0 : s et j = 0 : t

$M_{i,j} = \min (M_{i-1,\, j} + C_{insi} (S_j), M_{i,\, j-1} + C_{suppr} (T_i), M_{i-1,\, j-1} + C_{subst} (S_j, T_i))$

Distance - Levenshtein $= M_{s,j}$

Avec :

$M_{i,j}$: L'élement de la matrice M situé à la ligne i et à la colonne j.

S_j : La j^{ème} lettre du mot S.

T_i : La i^{ème} lettre du mot T.

Pour comprendre le fonctionnement de cet algorithme, on prend l'exemple suivant :

Exemple : Soient les chaînes de caractères S= 'NICHE' et T= 'CHIENS' de longueurs respectives s=5 et t=6. Intuitivement on peut transformer la chaîne S en chaîne T en 5 étapes :

- Suppression de N et I $\rightarrow$ CHE
- Insertion de I, N, et S $\rightarrow$ CHIENS

Alors, la distance de Levenshtein « d » entre les mots « S » et « T » est d'au plus 5, pour le vérifier formellement, les tableaux 1 et 2 ci-dessous illustrent les résultats de la première et la deuxième étape. On applique cet algorithme sur les mots « S » et « T ».

		C	H	I	E	N	S
	0	1	2	3	4	5	6
N	1	0	0	0	0	0	0
I	2	0	0	0	0	0	0
C	3	0	0	0	0	0	0
H	4	0	0	0	0	0	0
E	5	0	0	0	0	0	0

Tableau 1 : Première étape de cette Algorithme

		C	H	I	E	N	S
	0	1	2	3	4	5	6
N	1	1	2	3	4	4	5
I	2	2	2	2	3	4	5
C	3	2	3	3	3	4	5
H	4	3	2	3	4	4	5
E	5	4	3	3	3	4	5

Tableau 2 : Deuxième étape de cette Algorithme

Le résultat de l'application de la distance de Levenshtein entre les chaînes S et T se trouve dans le nombre désigné par M[s,t] .

3. Application lexicographique

Pour appliquer la méthode de Levenshtein, nous avons élaboré un corpus modeste de 120 mots qui sont lexicographiquement équivalents dans les trois variantes de la langue Tamazight : Tamazight, Tachelhit et Tarifit et nous avons aussi implémenté l'algorithme associé à cette distance pour mesurer les distances entre chaque couple de mots lexicographiquement équivalents dans chaque paire de variantes.

3.1. Corpus de travail

Pour être cohérent avec notre objectif de normalisation de la langue Tamazight, nous avons construit notre corpus de travail à partir des dictionnaires et ouvrages en usage à l'Institut Royal de la Culture Amazighe (IRCAM) (Cid Kaoui, 1907 ; Serhoual, 2002). La Figure 1 ci-dessous illustre un exemple de quelques mots du corpus.

La deuxième, troisième et quatrième colonne représentent respectivement la l'écriture en variantes Tarifit, Tachelhit et Tamazight tandis que la première colonne donne le mot en français.

Fr	Tarifit (A.P.I)	Tachelhit (A.P.I)	Tamazight(A.P.I)
Cœur	ⵓⵍ (ul)	ⵓⵍ (ul)	ⵓⵍ (ul)
Main	ⴼⵓⵙ (fus)	ⴰⴼⵓⵙ (afus)	ⴰⴼⵓⵙ (afus)
Chemin	ⴰⴱⵔⵉⴷ (abrid)	ⴰⵖⵔⴰⵙ (aγras)	ⴰⴱⵔⵉⴷ (ɑbṟid)
Étoile	ⵉⵜⵔⵉ (itṟi)	ⵉⵜⵔⵉ (itṟi)	ⵉⵛⵔⵉ (iʃṟi)
Viande	ⴰⵢⵙⵓⵎ (aysum)	ⵜⵉⴼⵢⵢⵉ (tifyyi)	ⴰⴳⵙⵓⵎ (agsum)
Soleil	ⵜⴼⵓⵛⵜ (ʧfuʃt)	ⵜⴰⴼⵓⴽⵜ (tafukt)	ⵜⴰⴼⵓⴽⵜ (tafukt)
Femme	ⵜⴰⵎⵖⴰⵔⵜ (tamγaṛt)	ⵜⴰⵎⵖⴰⵔⵜ (tamγaṛt)	ⵜⴰⵎⵟⵟⵓⵜ (tamṭṭut)
Homme	ⴰⵔⵢⴰⵣ (aṟyaz)	ⴰⵔⴳⴰⵣ (aṟgaz)	ⴰⵔⴳⴰⵣ (aṟgaz)
Ciel	ⴰⵊⵏⵏⴰ (aʒnna)	ⵉⴳⵏⵏⴰ (ignna)	ⵉⵊⵏⵏⴰ (iʒnna)
Ancien	ⴰⵇⴱⵓⵔ (aqbuṟ)	ⴰⵣⴰⵢⴽⵓ (azayku)	ⴰⵣⴰⵢⴽⵓ (azayku)
Bruit	ⵔⵃⴰⵙ (ṟḥas)	ⵍⵃⴰⵙ (lḥas)	ⵍⵃⴰⵙ (lḥas)
Canal	ⵜⴰⵔⵢⴰ (taṟya)	ⵜⴰⵔⴳⴰ (taṟga)	ⵜⴰⵔⴳⴰ (taṟga)

Figure 1 : Extrait du corpus

3.2. Résultats expérimentaux

Nous avons ensuite appliqué la distance de Levenshtein sur les mots écrits en caractères tifinaghes en tant qu'alphabet officiel adopté par le Maroc. En voici un exemple d'application sur deux mots :

Soient les mots écris en tifinaghe X = 'ⵜⴰⵎⵖⴰⵔⵜ' Y= 'ⵜⴰⵎⴰⵟⵟⵓⵜ' correspondants le mot 'Femme' en français, ces mots représentent respectivement les équivalents en Tachelhit et Tamazight (du Moyen Atlas). La distance de Levenshtein entre X et Y est égale à 4.

Le tableau 1 ci-dessous montre comment est fait le calcul de la distance par le biais de la programmation dynamique. La distance de Levenshtein entre les deux mots symbolisés par les chaînes X et Y s'affecte au nombre M[s,t] qui s'actualise à chaque étape du calcul de cette distance.

La suite des opérations correspond à :

- la substitution des lettres (caractères) ⵜ, ⵔ et ⵎ par eux-mêmes,
- la suppression des caractères ⵖ, ⴰ et ⵔ,
- puis l'insertion de ⵟ, ⵟ et ⵓ, et enfin une substitution de ⵜ par ⵜ.

		ⵜ	ⴰ	ⵎ	ⵟ	ⵟ	ⵓ	ⵜ
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ⵜ	**1**	0	1	2	3	4	5	6
ⴰ	**2**	1	0	1	2	3	4	5
ⵎ	**3**	2	1	0	1	2	3	4
ⵖ	**4**	3	2	1	1	2	3	4
ⴰ	**5**	4	3	2	2	2	3	4
ⵔ	**6**	5	4	3	3	3	3	4
ⵜ	**7**	6	5	4	4	4	4	3

Tableau 3 : Exemple de calcul de la distance de Levenshtein

Nous avons ensuite appliqué la méthode sur tout le corps du travail composé de 120 mots comme illustré dans le tableau suivant :

Mots	Tarifit	Tachalhit	Tamazight	D1	D2	D3
Puits	ⴰⵏⵓ	ⴰⵏⵓ	ⴰⵍⵃⴰⵙⵉ	0	5	5
Cœur	ⵓⵍ	ⵓⵍ	ⵓⵍ	0	0	0
chemin	ⴰⴱⵔⵉⴷ	ⴰⵖⵔⴰⵙ	ⴰⴱⵔⵉⴷ	3	0	3
Viande	ⴰⵢⵙⵓⵎ	ⵜⵉⴼⵢⵢⵉ	ⴰⴳⵙⵓⵎ	6	1	6
Soleil	ⵜⴼⵓⵛⵜ	ⵜⴰⴼⵓⴽⵜ	ⵜⴰⴼⵓⴽⵜ	2	2	0
Femme	ⵜⴰⵎⵖⴰⵔⵜ	ⵜⴰⵎⵖⴰⵔⵜ	ⵜⴰⵎⵟⵟⵓⵜ	0	3	3
homme	ⴰⵔⵢⴰⵣ	ⴰⵔⴳⴰⵣ	ⴰⵔⴳⴰⵣ	1	1	0
Toujours	ⴽⵓⵙⵙ	ⴽⵓⴰⵙ	ⴽⵓⵙⵙ	1	0	1
Ciel	ⴰⵊⵏⵏⴰ	ⵉⴳⵏⵏⴰ	ⵉⵊⵏⵏⴰ	2	1	1
Ancien	ⴰⵇⴱ ⵓⵔ	ⴰⵣⴰⵢⴽⵓ	ⴰⵣⴰⵢⴽⵓ	5	5	0
Coq	ⵢⴰⵥⵉⴹ	ⵢⴰⵥⵉⴹ	ⴰⴼⵔⵔⵓⵊ	0	6	6
éclairage	ⵜⴼⴰⵡⵜ	ⵜⵓⴼⴰⵡⵜ	ⴰⵙⵉⴷ	1	5	6
Fille	ⵜⴰⵃⵏⵊⴰⵔⵜ	ⵜⴰⴼⵔⵓⵅⵜ	ⵜⴰⵔⴱⴰⵜ	5	4	3
Idiot	ⴰⴱⵓⵀⴰⵍⵉ	ⴰⴱⵀⵍⵓⵍ	ⴰⵎⴱⵀⵍⵓⵍ	4	5	1

Tableau 4: La distance de Levenshtein entre certains mots équivalents

D1 : la distance de levenshtein entre la variante Tarifit et Tachalhit.

D2: la distance de levenshtein entre la variante Tarifit et Tamazight.

D3 : la distance de levenshtein entre la variante Tamazight et Tachalhit.

Les résultats obtenus par l'application de distance de Levenshtein sur notre corpus de taille 120 mots homonymes de la langue Tamazight, montrent que la similarité est très forte entre ces mots vu qu'elle n'excède pas le nombre 6 si l'on calcule la distance deux à deux : c'est-à-dire entre chaque paire de mots équivalents dans deux variantes différentes.

Mais pour avoir une idée bien précise de l'équivalence entre les différentes variantes d'un mot, on propose d'utiliser une autre distance calculée à base des distances utilisées précédemment. Le besoin de définir un nombre qui renseigne la distance globale entre les différentes variantes s'impose donc. Plusieurs idées nous ont survenu pendant notre travail: La moyenne arithmétique, la moyenne quadratique ou le maximum des distances. Nous avons choisi de travailler d'abord par les deux premières, c'est-à-dire utilisant la moyenne arithmétique et la moyenne quadratique.

Le tableau 5 montre un exemple d'utilisation de ces moyennes ou distances généralisées de Levenshtein appliqués à quelques mots équivalents.

Mot	Moyenne Arithmétique	Moyenne Quadratique
Puits	3 ,33	4.08
Cœur	0	0
Chemin	2	2,44
Viande	4,33	4,93
Soleil	1,33	1,63
Femme	2	2,44
Toujours	0,66	0,81
Ciel	1,33	1,41
Ancien	3,33	4,08
Idiot	3,33	3,74

Tableau 5 : Les moyennes des distances entre quelques mots équivalents

Compte tenu de ces nombres que nous avons associés aux distances entre les différentes équivalences d’un même mot dans les 3 trois variantes de la langue Tamazight, il ressort que :

- l’équivalence doit être étudiée avec précision tout en tenant compte de la dialectisation de Tamazight. Ce qui veut dire encore qu’il est impératif de réduire ces distances en évoquant quelques équivalences phonétiques entre les syllabes ou les phonèmes qui causent la différentiation d’écriture d’un mot.
- Il ne faut pas chercher l’équivalence « à tout prix » car parfois le mot dont l’équivalence est à prouver ou à établir pourra tout simplement être considéré comme un synonyme quel que soit la façon de l’écrire.

4. Conclusion

Nous avons présenté dans cet article une étude de l’équivalence lexicographique entre les 3 variantes de la Langue Tamazight en utilisant la distance de Levenshtein. Nous envisageons dans la suite de notre travail d’utiliser d’autres distances pour comparer les mesures d’équivalence et d’élargir notre corpus d’étude pour essayer de formuler plus d’assertions et d’hypothèses sur les équivalences lexicographiques et phonétiques qui aideront sans aucun doute à l’accélération du processus de normalisation de la langue Tamazight.

5. Références

Ameur M., Boumalk A. (2004). Standardisation de l’amazighe, Rabat : IRCAM.

Cid Kaoui S. (1907). Dictionnaire Français-Tachalh’it et Tamazir’t (Dialectes berbères du Maroc). Paris : Ernest Leroux.

Elsa N. (2013). Comparaison de textes : quelques approches. Rapport interne n°338, Laboratoire d’analyse et modélisation des systèmes pour l’aide à la décision, Université DAUPHINE. Paris, pp. 6-7.

Gueddah H., Yousfi A., Belkasmi M. (2012). Introduction of the weight edition errors in the Levenshtein distance. *International Journal of Advanced Research in Artificial Intelligence*, 1(5):30-32.

Kukich K. (1992). Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377-439.

Levenshtein V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *SOL Phys Dokl*, pp. 707-710.

Serhoual M. (2002). Dictionnaire Tarifit-Français. Thèse de Doctorat, Faculté des lettres et des sciences humaines, Téouan.

Wechsler M., Munteanu E., Schauble P. (1998). New techniques for open-vocabulary spoken document retrieval. *SIGIR*, pp. 20-27.

TICAM'14, p. 81-91

Towards an Amazigh Digital Dictionary and a Spell Checker

Mustapha Eddahibi Ali Rachidi
Research team TISED, National School of Business and Management,
Agadir, Morocco
{m.eddahibi, a.rachidi}@uiz.ac.ma

Abstract

This paper presents an Amazigh text editor that allows performing all basic text edition operations in addition to a spell-checking engine. It includes an Amazigh digital dictionary that provides words synonyms and translations. The digital dictionary is an open source application that uses the python dictionary data structure. To fill the lack in term of Amazigh corpus, the dictionary could be of enormous use in natural language processing field. It can be used successfully in many applications, say for example spellchecker that we have developed in the same application. This spellchecker is also an open source application that checks the existence of each word in the dictionary and suggests more likely substituents.

1. Introduction

The use of new information and communication technologies (NICT) has become essential for the survival of heritage in society. Indeed, the development of tools, for processing and analyzing Amazigh text, can provide significant support for the Amazigh language and a real chance to survive on the web and in the computerized world.

The application of these new technologies will highlight the Amazigh language phenomena. Indeed, we can do word processing and e-mails or even search or use the syntactic and grammatical features of the language. But, it is not the case of Amazigh language that lacks support in NICT environments. Therefore, scientific and linguistic research is engaged to remedy this situation. One of the priority components of this research is to design and build applications that can automatically process Amazigh language data.

In addition to the Amazigh text editor provided with the basic operations, this paper presents also a toolbar that loads automatically in MS Office applications. The toolbar configures the system for the Amazigh text composition.

The rest of the paper is organized as follows: in the next section, we present the Amazigh lexicon and its linguistic characteristics. Then, we present the text editor that we have developed and implemented. The immediately following section will present the digital dictionary. In the next section, we present the spell checking process. In the conclusion, we discuss different strength and weakness of these tools to prepare for the launch of future research works.

2. Amazigh language

2.1. Presentation

The Amazigh language is the oldest attested language in the Maghreb. Its area covers nearly five million km². It extends from east to west, from the Egyptian-Libyan borders to the Canary Islands, and from north to south from the southern shore of the Mediterranean to Niger, Mali and Burkina Faso. The most important community whose first language is Amazigh is located in Morocco. By its anteriority, the Amazigh language is the expression mode of the first identity of Moroccans. It is an essential foundation of their socio-cultural environment as it shapes their collective unconscious and marks their basic personality. It currently plays important role in Amazigh cultural movement.

2.2 Amazigh Alphabet

IRCAM has proposed tifinagh alphabet to the International Standardization Organisation. This proposal includes four subsets of tifinagh characters:

- IRCAM basic set;
- IRCAM extended set;
- Other neotifinagh letters;
- The modern Touareg letters whose use is attested.

Tifinagh alphabet and the Unicode associated plane assigned by ISO is shown in Figure 1.

2.3. Punctuation

We do not know if there exists particular tifinagh punctuation. IRCAM has recommended the use of conventional signs that are found in Latin-based scripts: “ ” (space), “.” , “,” , “;” , “:” , “?” , “!”, etc. Consequently, this recommendation doesn’t contain any tifinagh punctuation sign.

2.4. Sorting

Only IRCAM has defined a specific order described by the following expression (a <b means that a is sorted before b):

ⴰ < ⴱ < ⴳ < ⴳⵯ < ⴷ < ⴹ< ⴻ < ⴼ < ⴽ < ⴽⵯ < ⵀ

ⵀ < ⵃ < ⵄ < ⵅ < ⵇ < ⵉ < ⵊ < ⵍ < ⵎ < ⵏ < ⵓ < ⵔ

ⵔ < ⵕ < ⵖ < ⵙ < ⵚ < ⵛ < ⵜ < ⵟ < ⵡ < ⵢ < ⵣ < ⵥ

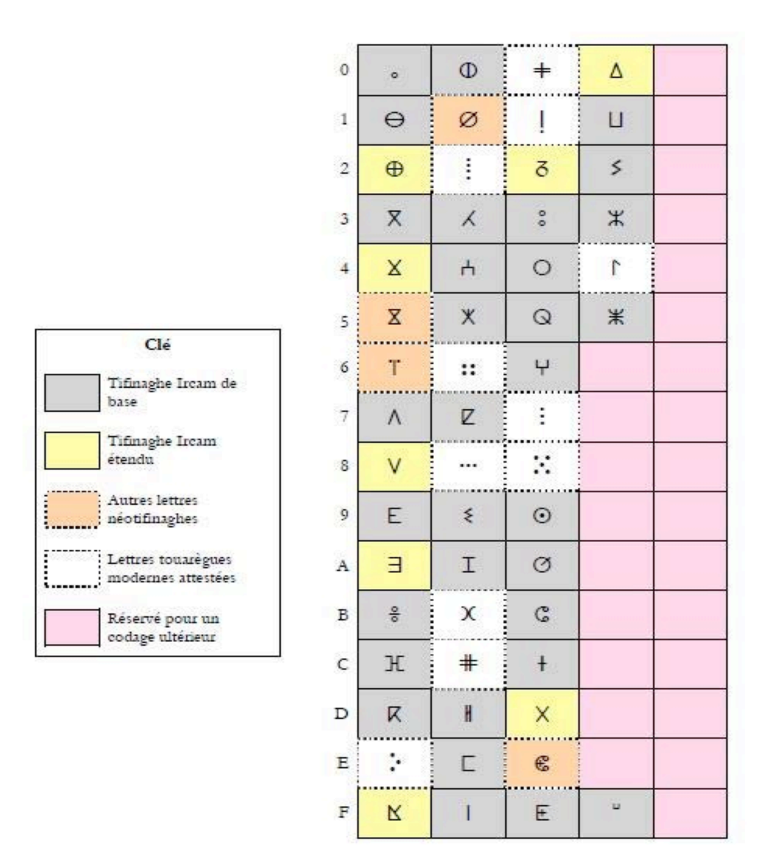


Figure 1: Tifinagh associated Unicode

2.5. Digits

IRCAM has retained “Arabic” Western digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) for tifinagh script. This proposal does not introduce any new digit or number.

2.6. Directionality

IRCAM has retained horizontal direction from left to right for tifinagh script.

2.7. Amazigh keyboard

Tifinagh alphabet contains thirty three characters. Figure 2 shows a keyboard in ASCII format proposed by the CEISIC center, at IRCAM (IRCAM, 2003 a; IRCAM, 2003 b).

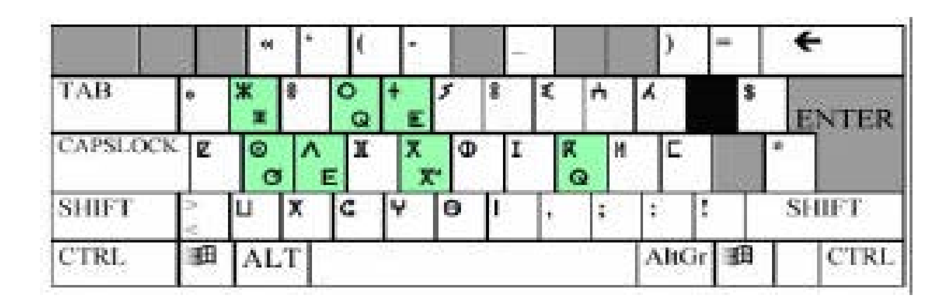


Figure 2: Tifinagh keyboard in ASCII format, the 26 first characters are directly accessible.

3. Amazigh lexicon organizing principles

Amazigh language has a very rich and a very complex morphology. There are three forms of Amazigh words: noun, verb and particle (Boukhris, 2008). The word classification in a dictionary can be done, basically, in two ways:

- Words alphabetical classification based on root.
- Words alphabetical classification based on derivatives words.

To develop an Amzigh dictionary, some difficulties will be faced. Namely:

- Verbs spelling change according to their conjugation.
- Nouns spelling change according to number (Passing from singular to plural).

The question then arise as to know whether there is a need to insert in the eventual dictionary, all nouns and verbs words. The answer is obviously, no. What should be done, in our opinion, is to stabilize verbs and nouns, in their variations respectively according to conjugation and in number.

The idea will be then, attempting to reduce variations as much as possible, without getting too far from the phonological form, in order to avoid reaching a writing that will turn away Amazigh learners (being Amazighophone or not). The spelling of same family words should refer to the base word (lemma), which is present in the Amazigh dictionary.

In addition, the spelling is like a utility that allows representing in writing form an exhaustive liable set of lingual information element belonging to different components of the Amazigh lexicon and grammar (phonology, morphology, syntax). Its aim is to render a text without any contextualization, the most explicit and the least ambiguous.

To be able to create the current system, we have proposed a dictionary of Amazigh terms.

4. Design and implementation of an Amazigh text editor

The Amazigh language lacks support in computerized environments. Indeed, the implementation of an Amazigh text editor, independent and autonomous is an important step in the computerization of the Amazigh language. The first part of our work is to make a text editor of the Amazigh language (Ait zegui *et al*., 2012). The idea of designing a text editor for the Amazigh language came from the fact that the text editors available on the word processing market does not offer all the features needed for a word processing quality. The text editor that tries to achieve ensures the basic features of word processing such as:

- Open File
- Entering Amazigh text
- Canonicalization selected text (standardization, entry not unique)
- Copy and paste Amazigh text
- Change the font
- Save the file in RTF

Functionality that deal more to this editor and search function and replacement of Amazigh text, as this is not taken into account for the Amazigh language on text editors already.

5. Developed digital Amazigh dictionary

Digital dictionary is based on a text file that is formatted as inputs (lines). Each input contains an Amazigh term and a set of related terms. Related terms form a list of words separated by the character "-". There are two categories of Amazigh terms in our dictionary: nouns and verbs.

- If the input is a noun , its associated terms are: its annexation state , its plural, its Latin-based transliterations and translations in French, English and Arabic;
- For verbs, related terms are the four aspects (aorist, perfective, negative perfective, imperfective) its Latin-based transliterations and translations in French, English and Arabic.

This dictionary was created using text extraction and reorganization techniques applied to a pdf file of the book "amawal n tjerrumt" published by the IRCAM in 2009 (IRCAM 2009).

Text extraction was done using the package "pdfminer" which is a Python library. Thanks to this package, we have been able to extract text of the pdf file from page 22 to page 59. The choice of the pages is justified by the fact that the text of each input line in the dictionary was written in three separated lines see (Figure 1):

- The first line contains translations;
- The second gives words forms in Amazigh;
- The third corresponds to second lines Amazigh words transliterations.

• **Terme / Term / مصطلح**

ⵉⵔⵎ – ⵢⵉⵔⵎ – ⵉⵔⵎⴰⵡⵏ

irem - yirem - irmawen

Figure 1: Lines of text extracted from the books "amawal n tjerrumt"

After text extraction, one part of the preprocessing code was dedicated to the reorganization and purification of the text lines that forms the dictionary input lines using the library "re" that processes regular expressions. Figure 2 shows the resulting digital dictionary in a TXT format.

dictionnaire.txt - Bloc-notes

Fichier Edition Format Affichage ?

ⴰⴽⵓⴷ - ⵡⴰⴽⵓⴷ - akud - wakud - Temps - Tense - زمن
ⵓⴷⴷⵉⵣ - ⵡⵓⴷⴷⵉⵣ - ⵓⴷⴷⵉⵣⵏ - uddiz - wuddiz - uddizen - Tendu - Tense - مشدد
ⵓⵙⵙⵉⴷ - ⵡⵓⵙⵙⵉⴷ - ⵓⵙⵙⵉⴷⵏ - ussid - wussid - ussiden - Tendu - Tense - مشدد
ⵜⵓⵙⵙⴷⴰ - ⵜⵓⵙⵙⴷⴰ - tussda - tussda - Tension - Tension - تشديد
ⵓⴷⵓⵣ - ⵡⵓⴷⵓⵣ - uduz - wuduz - Tension - Tension - تشديد
ⵉⵔⵎ - ⵢⵉⵔⵎ - ⵉⵔⵎⴰⵡⵏ - irem - yirem - irmawen - Terme - Term -مصطلح
ⵜⴰⴳⴰⵔⴰ - ⵜⴳⴰⵔⴰ - ⵜⵉⴳⴰⵔⵉⵡⵉⵏ - tagara - tgara - tigariwin - Terminaison - Ending - وقف
ⵜⴰⵙⵏⵉⵔⵎⵜ - ⵜⵙⵏⵉⵔⵎⵜ - tasniremt - tesniremt - Terminologie - Terminology - مصطلحيات
ⴰⴹⵕⵉⵙ - ⵓⴹⵕⵉⵙ - ⵉⴹⵕⵉⵙⵏ - aÄris - weÄris - iÄrisen - Texte - Text - نص
ⴰⵙⴳⵓⵎ - ⵓⵙⴳⵓⵎ - ⵉⵙⴳⴰⵎ - asgum - wesgum - isgam - Thème - Theme - محور
ⵜⴰⵡⵉⵍⴰ - ⵜⵡⵉⵍⴰ - ⵜⵉⵡⴰⵍ - tawila - twila - tiwal - Thème - Theme - مبنى
ⴰⵙⵓⵖⵍ - ⵓⵙⵓⵖⵍ - asuvel - usuvel - Traduction - Translation - ترجمة

Figure 2: Dictionary text file

The resulting text file is used by the spell checking module written in Python. The file content is stored in a Python data structure called dictionary. This is an extremely powerful and practical type of data. The Python dictionary is different from the list structure on several points. Lists are objects containers. To access an element in a list, one should know its position in the list, called also index. Python dictionaries are also containers. However, it doesn't use any ordered structure. Moreover, to access a dictionary element, it is not necessary to use indices but only keys. In our case, the first word of each line in the text file (Figure 2) is considered as a key and the other terms as values. See Figure 3 that shows our digital dictionary form on Python.

Dicti = { "ⵓⴷⴷⵉⵣ" : ["ⵡⵓⴷⴷⵉⵣ", "ⵓⴷⴷⵉⵣⵏ", "uddiz", "wuddiz", "uddizen" , "Tendu", "Tense", " مشدّد "],

……

}

Figure 3: An element of our digital dictionary

6. Lexical analysis in the developed text editor

6.1. Developed text editor

The conception of a text editor for Amazigh language was a need, knowing that the existing text editors and processing tools do not provide all the needed functionalities for a good Amazigh text processing.

The text editor that we are attempting to develop allows ensuring basic functionalities of the text processing such as:

- Opening files;
- Writing Amazigh text;
- Canonical layout of a selected text (standardization, unequivocal input);
- Amazigh text copy past;
- Changing font;
- Saving the file on a RTF format;
- Spell checking.

The functionalities that one has focused on for this text editor was search and replace functionalities for Amazigh text. Because, these operations aren't taken into account for Amazigh text, in any of the existed text editors.

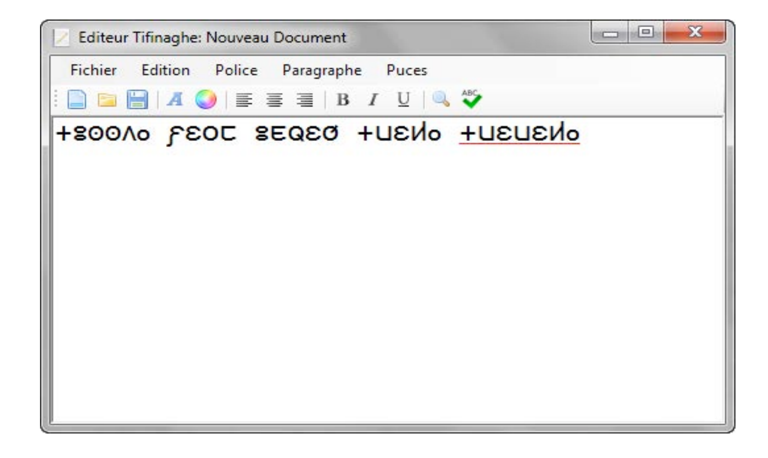


Figure 4: Graphical interface of the developed text editor with a built-in spell checker

7. Spell checking

The first spell checkers were widely available on mainframe computers in the late 1970s. For Amazigh language, the first spell checker was an adapted version of Hunspell (Y. Essaady *et al*., 2009). In this work, the spell checking operation is optional in the Amazigh text editor (Figure 4). Indeed, this functionality would be enabled only if the user clicks on the spell checking button in the tool bar. Once this option is enabled, all the words that aren't included in the dictionary text file, even if they aren't mistakes, they will be underlined in red color. The spell checking is done gradually for the last words entered by the user. The spell checking process is launched for the last word once the user tapes blank space key or moves the cursor.

The spell checking code is written in Python even if the graphical interface of the text editor was developed using CSharp in the Visual Studio IDE. The reason behind the choice of Python is its efficiency, its ease and its rapidity in text processing. Combining ".net" and Python can be done by using the library "ironPyton" that allows the interpretation of Python code. This solution isn't optimal because it is less fast and not scalable. The most efficient approach is the process approach where the Python code is used as string to be executed by the installed version of the Python interpreter. This solution gives greater freedom for using different Python versions and benefitting from different updates.

Spell checking is to verify whether the key which the researched Amazigh word exists or now in the Python dictionary. The code for the spell checking is reduced to a simple test whose code is as follows in Figure 5:

```
ifdict['mot_Amazigh'] != None:
errOrtho =true
```

Figure 5: Code checking the existence of a word in the dictionary

The dictionary provides other uses like transliteration and translation.

8. Amazigh toolbar for MS Office

4.1. Presentation

Microsoft Office is an office suite developed by Microsoft for Windows and Macintosh platforms. It includes several software tools. Microsoft Word is word processing software that is the central program of Microsoft Office. It dominates the market of word processing software.

The integration of the Amazigh language in Microsoft Office is a point in favor of the language. It will enable reliable and high quality processing of Amazigh text. For this reason, a part of our work was to achieve a Microsoft Office toolbar that configures the system for Amazigh text processing.

4.2. Implementation and involved tools

To achieve the MS Office toolbar, the most suitable way was to use the integrated development environment Microsoft Visual Studio 2010. Indeed, it includes new features that simplify application development task. Among these features it includes the development of Add-ins for MS Office.

An Add-in is a complementary part for software. It can be in the form of a main menu, a contextual menu or even a background function. Here, we will develop a menu to integrate features that are missing for Amazigh text processing by the MS Office suite.

Features that we are trying to integrate in MS Office are mainly sorting, Amazigh text search and replace, and Amazigh dictionary.

This menu that will be called AMAZIGH would be as follows:

Figure 5: AMAZIGH MS Office menu

From left to right, buttons functions are:

- Search and replace button will allow Amazigh text search and replace ;
- Font button will allow to use an Amazigh font ;
- Sorting button will sort words in a table. The sorting is performed using lexicographical order ;
- Dictionary button gives access to the Amazigh dictionary.

Add-in programming is done using the C sharp language.

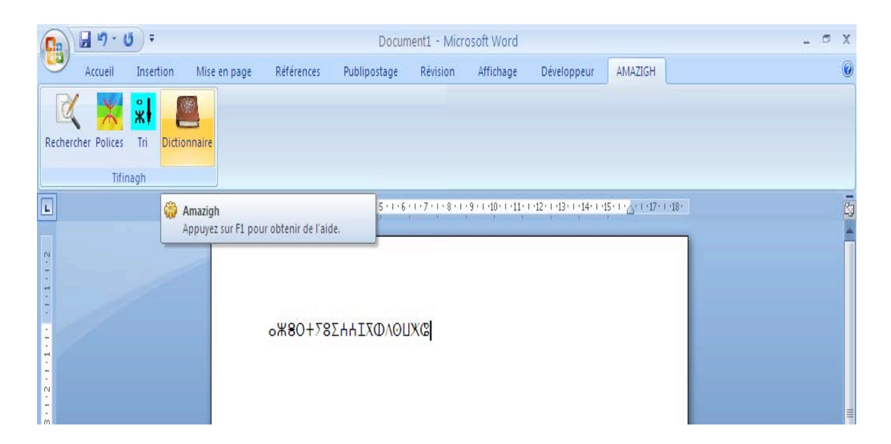


Figure 6: AMAZIGH Add-in in action

The improvement of Amazigh language support in Ms Office is currently ongoing.

8. Exploration, simulation and discussion

The term's file was used by the developed editor to check whether a word exists or not in the dictionary. We met some difficulties at the level of Amazigh morphology to create a file that includes all affixation rules of this language. It is often manually created and it is a hard work. It requires a lot of time, but it didn't require any computer skill. It is linguists' work because they know spelling and grammatical rules for Amazigh language.

This research work allows the initiation of a professional Amazigh editor development. Thus editor will be among main Amazigh language processing tools. However, it has some limits:

- The absence of a real standard Amazigh language dictionary, this hinders our editor and our dictionary.
- The absence of rich Amazigh corpus to enable spell checking tool to propose most likely words once a word didn't exist in the dictionary.
- The absence of exhaustive morpho-syntactic standard rules makes our editors very weak in the syntactic and lexical analyzis level.

While waiting for IRCAM and the center of language planning (CAL) to develop these linguistic tools, we are going to improve our editor technically in order to adapt new tools and trends of the Amazigh scientific community.

9. Conclusion

We have presented, in this paper, an Amazigh text processing system. The system provides basic functionalities and spell checking. This checking is based on a digital dictionary that we have developed on Python and an open source system that checks for every word of a taped text its existence in the dictionary. This system is based on an Amazigh corpus to propose words near to the mistake word using the concept of the most probable and the most possible. These work results were very successful and offer many different perspectives for text editor improvement by adding other known functionalities. This paper presents also an Amazigh toolbar for MS Office applications.

References

Centre des études informatiques et des systèmes d'information et de communication (2003). Élaboration d'une première version du clavier amazighe, Centre des études informatiques et des systèmes d'information et de communication, Plan d'action 2003, IRCAM, Rabat.

Centre des études informatiques et des systèmes d'information et de communication (2003). Conception et mise au point des polices Amazighes, Plan d'action 2003, IRCAM, Rabat.

Ait Zengui N., Rachidi A., Bouyakhf E. (2012). Réalisation d'un éditeur de texte pour la langue Amazighe et d'une barre d'outils Amazighe pour MS Office, *TICAM'12*, 26-27 Novembre 2012, IRCAM, Rabat, Maroc.

Boukhris F., Boumalek A., Souifi E. (2008). La nouvelle grammaire de l'Amazighe, IRCAM, Rabat, Maroc.

Boumalk A., Naït-Zerrad K. (2009). Vocabulaire grammatical, IRCAM, Rabat, Maroc.

EsSaady Y., Ait Ouguengay Y., Rachidi A., Elyassa M., Mammass D. (2009). Adaptation d'un correcteur orthographique existant à la langue Amazighe: cas du correcteur Hunspell, *Actes du 1er symposium international sur le traitement automatique de la culture Amazighe*, SITACAM, Agadir, Maroc.

TICAM'14, p. 93-100

Noms et adjectifs dérivés en Tamazight (Kabyle)

Farida Aoughlis
Université Mouloud Mammeri Tizi Ouzou Algérie
fariyamo@yahoo.fr

Résumé

Chaque verbe est susceptible de fournir des noms verbaux et des adjectifs par dérivation. Les noms verbaux sont le nom d'action, le nom concret, le nom d'agent, le nom d'instrument et l'adjectif. Le nom d'action existe pour tous les verbes contrairement aux autres noms. Dans le cadre du nouveau module Tamazight (kabyle) pour NooJ (Aoughlis, 2012 ; Aoughlis *et al.*, 2013), nous nous intéressons aux noms et adjectifs dérivés à partir des verbes du dictionnaire électronique existant (Aoughlis *et al.*, 2013). Nous décrirons dans ce papier les différents dérivés nominaux du verbe pour Tamazight (Kabyle) ainsi que les ressources linguistiques construites (règles de flexion / dérivation) à l'aide du système NooJ[1] en vue du TAL de Tamazight (kabyle). NooJ est une plate-forme de développement permettant la description des langues naturelles ou techniques, elle est d'usage libre. Une vingtaine de langues sont déjà disponibles, grâce à nos travaux actuels, Tamazight (kabyle) sera intégréé à NooJ.

1. Introduction

La dérivation (Chaker, 1995) se définit en linguistique générale comme la procédure de formation de mots par combinaison d'un élément lexical (appartenant à un inventaire ouvert) et d'un morphème grammatical (appartenant à un inventaire fermé). « La dérivation joue un rôle essentiel, tant dans la formation du lexique que dans la syntaxe de la phrase verbale. En principe, toute unité lexicale berbère est susceptible d'être décomposée en : 1° une racine lexicale consonantique (porteuse de la notion sémantique centrale) et, 2° un schème de dérivation déterminé, verbal ou nominal, qui affecte le complexe ainsi formé (« mot ») à une catégorie morpho-syntaxique particulière. Pour les nominaux, on identifie des schèmes de noms d'action, de noms d'agent, de noms d'instrument, d'adjectifs ... ».

Dans le cadre de nos travaux, les dictionnaires électroniques sont construits en vue de l'analyse automatique. Les types d'analyse morpho-lexicale (à base de racines et schèmes, à base de lexèmes et à base de lemmes) sont examinés. Une étude comparative est faite dans (Mesfar, 2008). Le lemme est la solution intermédiaire, c'est le type d'entrée retenue pour les dictionnaires construits dans NooJ. Les usagers qu'ils soient avertis ou non peuvent faire une recherche facile d'un mot de la langue dans ces dictionnaires. Dans ce papier, nous étudierons les différents noms et les adjectifs ainsi que leur mode de dérivation à partir des verbes. Dans

1 http://www.nooj4nlp.net/

le système NooJ, nous décrirons les règles de flexion et dérivation correspondant aux modes de dérivation définis dans les travaux de (Nait-Zerrad, 1995-b). Nous associerons au verbe le paradigme de dérivation qui lui est associé. Après la flexion du dictionnaire, nous récupérons les noms et adjectifs dérivés.

2. Le nom d'action

Le nom d'action signifie réaliser ou subir l'action exprimée par le verbe, par exemple :
ffer → tuffra *action de cacher*.

2.1. Modes de formation

Sept modes de formation des noms d'action qui sont les plus réguliers sont recensés dans (Nait-Zerrad, 1995-b). Le mode de formation I est le plus commun, il est employé avec des verbes de 2 à 3 syllabes. La formation du nom est faite en préfixant la voyelle initiale "a" directement au verbe. Si le verbe commence ou finit avec 2 consonnes identiques une seule subsiste (bberzegzew → berzegzew). Les modes de formation II, III et IV associent en plus de la préfixation, la suffixation d'une voyelle u ou i de la voyelle initiale "a". Les verbes concernés n'ont qu'une syllabe. Le mode V est le mode de formation pour les verbes trilitères et particulièrement les verbes de type 1. Le mode VI concerne les verbes tels que ffer *cacher*. Le mode VII est utilisé pour les verbes de qualité.

2.2. Modes, verbe et nom d'action

Les 7 modes de formation des noms d'action sont donnés dans le Tableau1 :

Mode	Forme du verbe	Nom d'action
I	$C_1V_1\ C_2V_2$	a $C_1V_1\ C_2V_2$
II	$C_1(C_2)\ V_1\ C_3\ (C_4)$	a $C_1(C_2)\ V_1\ C_3\ (C_4)\ V_1$
II	$C_1(C_2)$ e $C_3\ C_4$	a $C_1(C_2)$ e $C_3\ C_4$ i
IV	$(C_1\ )\ C_2$ a $C_3\ (C_4)$	a $C_1(C_2)$ a $C_3\ (C_4)$ i
V	$C_1\ C_2$ e C_3	a $C_1\ C_2$ a C_3
VI	$C_1\ C_2$ e C_3	t u $C_1\ C_2\ C_3$ a
VII	i $C_1\ C_2\ V_1\ C_3$	t e $C_1\ C_2$ e C_3

Tableau1 : Modes réguliers de formation du nom d'action [2]

2 C_i est une consonne a, i est la position de la consonne dans le mot, V_i est une voyelle, i est la position de la voyelle dans le mot.

2.3. Règles de flexion / dérivation

Dans le système NooJ (Silberztein, 2003), les grammaires flexionnelles et dérivationnelles sont représentées par des ensembles structurés de graphes ou de règles qui décrivent des paradigmes morphologiques stockés dans des fichiers ".nof".

Examinons dans le système NooJ la règle de formation du nom d'action, pour les types 1 et 8, nous avons les règles de formation V, I et Id:

N-ACTION5=<P>a<N><L2>a<S><N>/N-act5; **zdem → azdam**

signifie que "a" est rajouté au début du verbe, que "e" est remplacé par "a"

N-ACTION1= <P>a/N-act1; **friwes → afriwes**

signifie que "a" est rajouté au début du verbe,

N-ACTION1D= <P><S>a/N-act1d; **bberzegzew → aberzegzew**:

Signifie dans le cas particulier du mode I (si le verbe commence par 2 consonnes identiques), une consonne est supprimée et on rajoute au début du mot obtenu "a".

3. Le nom concret

Le nom d'action (Nait-Zerrad, 1995-b) peut avoir un sens concret, mais on peut trouver des noms concrets dérivés du verbe : (bri → abray *action de concasser et d'écraser*, bri → abruy *petit morceau* ; zdem → azdam *action de ramasser du bois*, zdem → tazdemt *fagot* ; gmer → agmer *action de cueillir*, gmer → tagmert *cueillette*).

4. Le nom d'agent

Il correspond à l'auteur du procès. Le nom d'agent (Nait-Zerrad, 1995-b) se forme par la préfixation de "am" au verbe, avec la variante "an" (non systématique) si le radical du verbe contient une labiale ("m", "b", "f", "w") : zuṛ *faire un pèlerinage* → amzuṛ *pèlerin*, gmer *cueillir* → anegmar *celui qui cueille*. Si le verbe contient la voyelle i, la voyelle initiale a devient i. La formation de ces noms dérivés varie selon les types de verbes.

Dans le cas du type 1, il y a apparition d'un "a" ou d'un "u" intraradical : gmer *cueillir* → anegmar *celui qui cueille*, zdeγ *habiter* → amezdaγ *habitant*, xleq créer → amexluq *créature*.

Dans le cas des types 63/64, il y a préfixation de "am" et apparition d'un "a" entre les deux radicales avec par exemple : agem *puiser* → anagam *celui qui puise*.

La règle de flexion NAGENT1= <P>ane<N><L2><S>a ; permet de générer le nom d'action anegmar à partir du verbe gmer.

La règle de flexion NAGENT2= <P>an<N><L2><S>a ; permet de générer le nom d'action anagam à partir du verbe agem.

Il y a aussi d'autres types : ṭṭef *tenir* /anaṭṭaf celui qui tient, niwel cuisiner / tamnawalt *cuisinière*, ẓed *moudre* / ameẓẓad meunier.

5. Le nom d'instrument

Ils sont rares en Kabyle. La formation du nom d'instrument se fait par suffixation d'un "s" au verbe. Certains sont confondus avec le nom d'action d'autres sont formés comme le nom d'agent. Rgel *bouchon* → asergel *bouchon*, Nqer → amenqar *burin*.

6. L'adjectif

L'adjectif exprime une propriété, une caractéristique (animé ou chose) ou une qualité du nom auquel il se rapporte. La dérivation se fait surtout à partir des verbes de qualité (d'état). Izwiγ *être rouge* → azeggaγ rouge mais d'autres adjectifs sont dérivés de verbes ordinaires. Zzelmeḍ *être à gauche* → azelmaḍ *de gauche*. On trouve des adjectifs formés sur le modèle du nom d'agent : llaẓ *avoir faim* → amellazu *affamé*.

6.1. Règles de formation à partir du verbe de qualité

L'adjectif spécifique est formé par suffixation de "an" verbe de qualité, pour les verbes, selon le nombre de consonnes radicales (mode 1) ou le prétérit (autres modes). La formation est décrite dans le Tableau 2 suivant :

Mode	Forme du verbe	adjectif	Exemple
1	$V_1C_1\ V_2C_2$	a $C_1V_2\ C_2$ an	Iẓid/aẓidan
2	$iC_1C_2\ iC_3$	a C_1 e $C_2\ C_3$ an	Ibrik/aberkan
3	$iC_1C_2 iC_3$	a C_1 e $C_2\ C_2$ a C_3	Ismiḍ/asemmaḍ
4	$iC_1C_2\ iC_3$	a $C_1\ C_2$ a C_3 an	Irẓig/arẓagan

Tableau 2 : Modes de formation à partir du verbe de qualité

6.2. Règles de formation à partir du verbe ordinaire

Pour les verbes ordinaires, l'adjectif se forme par préfixation de "am" ou de ses variantes au verbe, le procédé de suffixation de "an" existe aussi (qqar/aquran). D'autres procédés de formation existent (Nait-Zerrad, 1995-b).

7. Ressources linguistiques pour Tamazight (kabyle)

7.1. Le dictionnaire des verbes

Les verbes sont recensés à partir de (Nait-Zerrad,1995-a) puis codifiés dans NooJ. Par exemple dans l'entrée de dictionnaire de la Figure 1 suivante :

bges,V+FLX=krez+DRV=N-ACTION5:Flexion+FR=cintrer|sangler

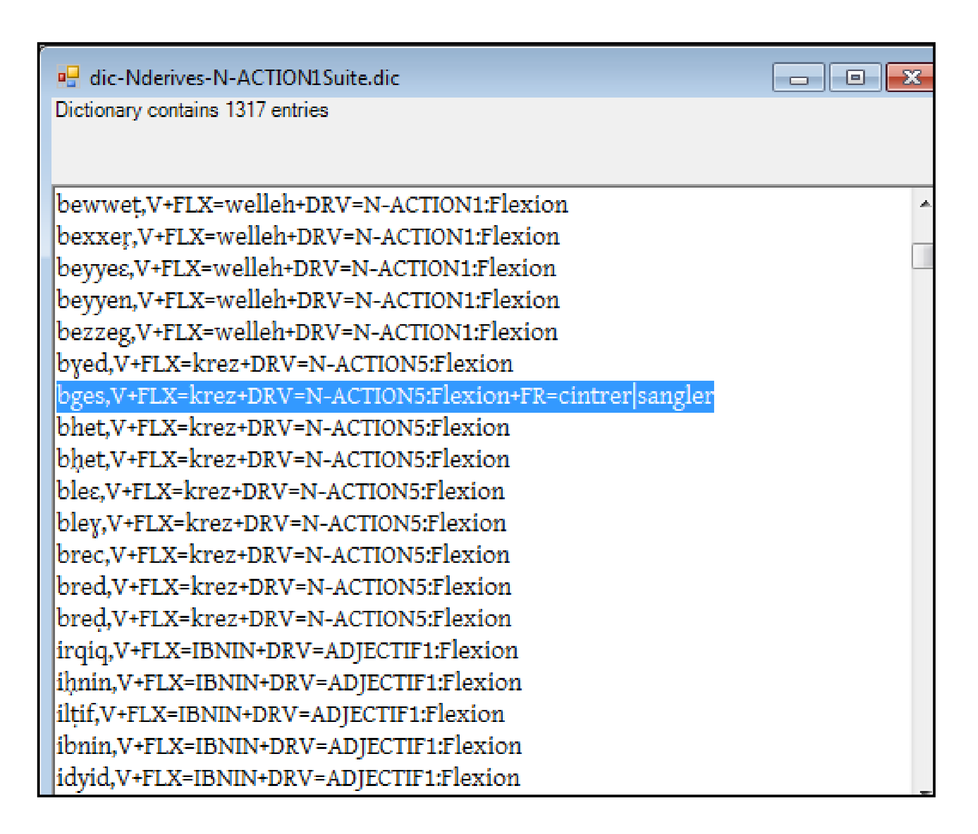


Figure1 : Extrait du dictionnaire des verbes utilisé

Le code "+FR" donne la traduction en français du verbe.

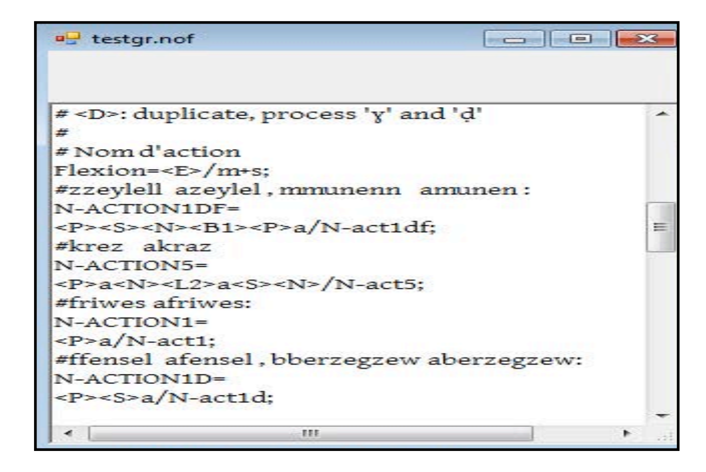
Le code "+FLX" (Paradigme flexionnel) permet d'indiquer le nom des règles de flexions du mot, bges est associé avec la propriété +FLX= krez, et Krez est un paradigme flexionnel.

Le code "+DRV" (Paradigme dérivationnel) permet aux lexicographes d'associer chaque mot avec les variantes dérivationnelles correspondantes. Des dérivations morphologiques seront ensuite décrites avec précision pour chaque entrée. Lier des formes dérivées dans le dictionnaire permettra d'avoir un dictionnaire moins volumineux. Dans DRV=N-ACTION5, N-ACTION5 est un paradigme dérivationnel qui permet de générer le nom d'action à partir du verbe krez. Dans le dictionnaire des formes fléchies des verbes on aura l'entrée abgas action de cintrer/sangler issue du verbe bges:

abgas,bges,Nact5+FLX=krez+DRV=NACTION5:Flexion+FR=cintrerl sangler+m+s

7.2. Les grammaires flexionnelles et dérivationnelles

On trouve dans les Figures 2 et 3, des extraits de la grammaire flexionnelle "testgr.nof" pour les noms d'action et les adjectifs :



```
testgr.nof
# <D>: duplicate, process 'ɣ' and 'ḍ'
#
# Nom d'action
Flexion=<E>/m+s;
#zzeylell azeylel , mmunenn amunen :
N-ACTION1DF=
<P><S><N><B1><P>a/N-act1df;
#krez akraz
N-ACTION5=
<P>a<N><L2>a<S><N>/N-act5;
#friwes afriwes:
N-ACTION1=
<P>a/N-act1;
#ffensel afensel , bberzegzew aberzegzew:
N-ACTION1D=
<P><S>a/N-act1d;
```

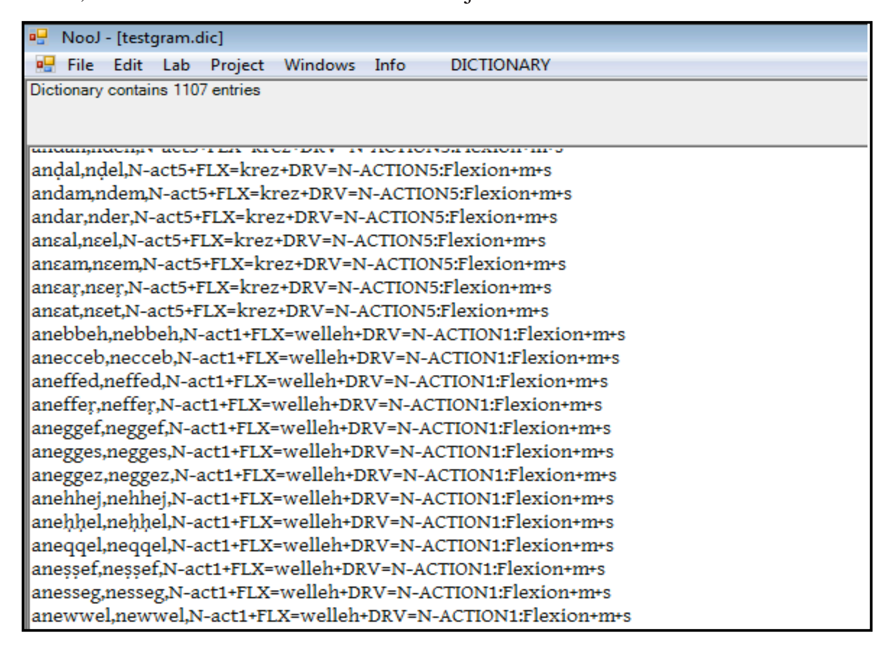
Figure 2 : Extrait de la grammaire flexionnelle/dérivationnelle : noms

```
#ADJECTIF-1
# iẓid aẓidan
ADJECTIF1=<P><S>a<N>an/Adjectif1;


#ADJECTIF-2
# ibrik aberkan
ADJECTIF2= <P><S>a<N>an<L5>e<R><S>/Adjectif2;

#ADJECTIF-3
# truẓi amtruẓi
ADJECTIF3= <P>am/Adjectif3;

#ADJECTIF-4
 #qqar aquran
ADJECTIF4= <P><S>a<N>an/Adjectif4;
```

Figure 3 : Extrait de la grammaire flexionnelle/dérivationnelle : adjectifs

8. Noms et adjectifs dérivés

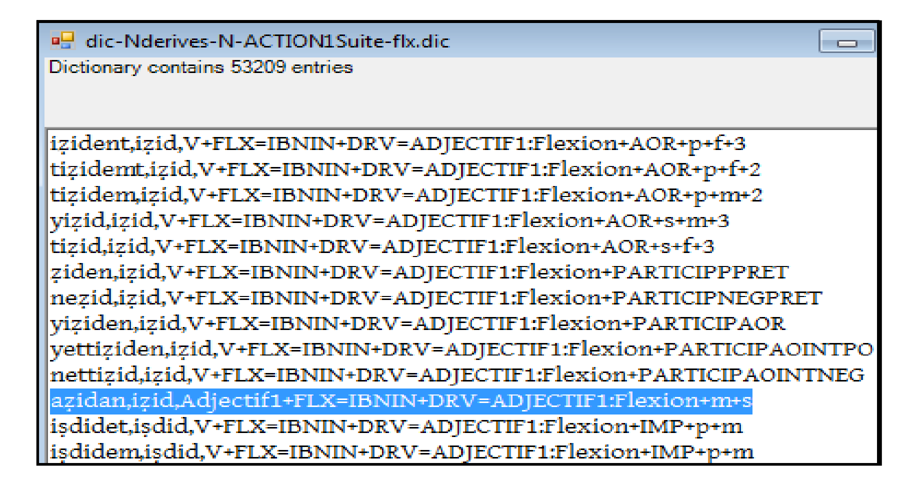
Ci- dessous, on trouve un extrait des noms et adjectifs dérivés :


NooJ - [testgram.dic]
File Edit Lab Project Windows Info DICTIONARY
Dictionary contains 1107 entries

anḍal,nḍel,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
andam,ndem,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
andar,nder,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
aneɛal,neɛel,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
aneɛam,neɛem,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
aneɛaṛ,neɛeṛ,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
aneɛat,neɛet,N-act5+FLX=krez+DRV=N-ACTION5:Flexion+m+s
anebbeh,nebbeh,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
anecceb,necceb,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneffed,neffed,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneffeṛ,neffeṛ,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneggef,neggef,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
anegges,negges,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneggez,neggez,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
anehhej,nehhej,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneḥḥel,neḥḥel,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneqqel,neqqel,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
aneṣṣef,neṣṣef,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
anesseg,nesseg,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s
anewwel,newwel,N-act1+FLX=welleh+DRV=N-ACTION1:Flexion+m+s

Figure 4 : Extrait du dictionnaire des noms et adjectifs dérivés

On notera que les noms et adjectifs dérivés sont obtenus après flexion du dictionnaire des verbes retenus. Nous donnons un extrait ci-dessous :


dic-Nderives-N-ACTION1Suite-flx.dic
Dictionary contains 53209 entries

iẓident,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+AOR+p+f+3
tiẓidemt,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+AOR+p+f+2
tiẓidem,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+AOR+p+m+2
yiẓid,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+AOR+s+m+3
tiẓid,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+AOR+s+f+3
ẓiden,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+PARTICIPPPRET
neẓid,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+PARTICIPNEGPRET
yiẓiden,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+PARTICIPAOR
yettiẓiden,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+PARTICIPAOINTPO
nettiẓid,iẓid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+PARTICIPAOINTNEG
aẓidan,iẓid,Adjectif1+FLX=IBNIN+DRV=ADJECTIF1:Flexion+m+s
iṣdidet,iṣdid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+IMP+p+m
iṣdidem,iṣdid,V+FLX=IBNIN+DRV=ADJECTIF1:Flexion+IMP+p+m

Figure 5 : Extrait du dictionnaire fléchi des verbes

9. Conclusion et perspectives

Dans ce travail, nous avons étudié et construit les ressources linguistiques pour la langue Tamazight (Kabyle) afin de dériver les noms et adjectifs à partir des verbes du dictionnaire déjà construit dans NooJ. En rajoutant un paradigme de dérivation à chaque entrée (lemme) des verbes fournissant des dérivés, la taille des dictionnaires est réduite, il n'y a pas lieu de mettre ces noms et adjectifs dans un dictionnaire, ils seront reconnus grâce aux paradigmes dérivationnels. Les ressources linguistiques (dictionnaires électroniques et grammaires) construites permettent le traitement automatique des langues (TAL), naturelles ou techniques, dans notre cas le Kabyle.

Le travail est en cours pour les autres types de verbes. Les exceptions et cas particuliers feront l'objet d'une codification comme entrées dans le dictionnaire des noms (en cours d'élaboration dans NooJ).

Références

Aoughlis F. (2012). Vers un module Tamazight pour le système NooJ. Iles d imesli, Vol. 4, pp. 229-244,Université Mouloud Mammeri,Tizi Ouzou.

Aoughlis F., Nait-Zerrad K., Annouz H., Aït-Kaci F., Habet MS. (2013). A new Tamazight Module for NooJ. *Formalising Natural Languages with NooJ 2013*. Koeva S., Mesfar S., Silberztein M. Eds. Cambridge Scholars Publishing: Newcastle upon Tyne.

Chaker S. (1995). Dérivation, *Daphnitae – Djado, Encyclopédie Berbère 15*, pp. 2277-2279. Edisud, Aix-en-Provence

Mesfar S. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. Thèse de Doctorat, Université de Franche-Comté, France.

Nait-Zerrad, K. (1995-a), Manuel de conjugaison kabyle, 6000 verbes, 176 conjugaisons (le verbe en berbère), Enag Ed. Alger.

Nait-Zerrad, K. (1995-b), *Grammaire du berbère contemporain (kabyle), I- Morphologie*, Enag Ed. : Alger

Silberztein M. (2003). Manuel NOOJ. http://www.nooj4nlp.net

TICAM'14, p. 101-110

Lexique amazighe pour mobile

Jamal Frain **Fadoua Ataa Allah** **Youssef Ait Ouguengay**
CEISIC, Institut Royal de la Culture Amazighe, Maroc
{frain, ataaallah, aitouguengay@ircam.ma}

Résumé

A travers cet article, nous focalisons notre attention sur la transmission et la diffusion de la langue amazighe via les technologies mobiles. A cette ère des nouvelles technologies de la communication et de l'information, le mobile constitue un canal d'information instantané et permanent, qu'il faut nécessairement exploiter pour promouvoir la langue amazighe et favoriser son apprentissage. Dans ce sens, notre contribution, consiste à présenter les spécificités de l'application mobile du lexique amazighe « LEXAM » publiée par l'IRCAM, en comparant cette dernière aux applications, existantes sur le marché, centrées sur le même sujet.

1. Introduction

De nos jours, le nombre d'utilisateurs de la technologie mobile ne cesse d'augmenter de manière impressionnante. Ce qui est justifié par les avantages qu'offrent les applications mobiles. D'ailleurs, le mobile est devenu le nouveau support du développement d'Internet. Il permet à l'utilisateur d'accéder à l'information recherchée tout le temps et à tout moment et d'utiliser facilement le mode téléchargement qui lui permet de consulter de façon autonome toute la documentation désirée.

Pareillement, les dictionnaires électroniques ont plusieurs avantages. Ils peuvent contenir un nombre important d'entrées lexicales et une variété d'information. Ils assurent via une simple recherche l'accès rapide à une multitude d'informations : définition, orthographe, prononciation, exemples et des locutions associés, synonymes et conjugaison de verbes. Ainsi, les dictionnaires électroniques permettent d'économiser un temps précieux de recherche et donnent la possibilité de prioriser la réflexion, le choix des idées, la construction de phrases et tous les aspects qu'aucun livre seul ne peut traiter à leur place. Les dictionnaires électroniques en font alors des outils d'acquisition du vocabulaire, d'enrichissement culturel et d'aide à l'écriture.

Après cette brève argumentation sur l'importance de la technologie mobile et des dictionnaires électroniques, nous pouvons dire que la combinaison de ces aspects favorisera, sans doute, la transmission et la diffusion de la langue amazighe et participera dans sa promotion. En commençant de ce point, nous pouvons souligner que l'objectif principal de cet article est de présenter une application mobile, qui est destinée non seulement aux amazighophones mais à tous ceux et celles qui s'intéressent à la langue amazighe, contenant un dictionnaire trilingue facile à utiliser même en mode déconnecté. Et nous nous concentrons par la suite sur les perspectives qui peuvent donner naissance à des ressources qui alimentent la réflexion sur la mobilité et la promotion de la langue amazighe.

2. Tifinaghe et les technologies mobiles

2.1. Historique du Tifinaghe

Bien que l'amazighe était principalement une langue de tradition orale, les amazighophones possèdent, depuis l'antiquité, leur propre système d'écriture appelé « Libyco-berbère » (le Tifinaghe en amazighe). Ce système est de nature alphabétique, à tendance phonologique fondé sur des signes à valeur consonantique, à usages traditionnellement assez restreints (funéraires, symboliques et ludiques). Il date de plus de 40 siècles (Hachid, 2001 ; Skounti *et al*., 2003). Cependant, les formats d'apparition de ses signes n'ont cessé de se développer : depuis son origine, le Libyque, jusqu'à le néo-tifinaghe, à la fin des années soixante, et le Tifinaghe-IRCAM, en 2001(Ameur *et al*., 2004). Historiquement, dans les anciennes inscriptions, l'amazighe s'écrivait dans toutes les directions et encore en boustrophédon (Lhote, 1979). Cependant l'orientation la plus souvent adoptée dans l'écriture amazighe moderne est horizontale et de gauche à droite.

Dans l'objectif de fournir à la langue amazighe un système alphabétique standard plus adéquat et utilisable pour tous les parlers amazighs actuels du Maroc, l'IRCAM a préconisé les caractères tifinaghes (ⴰ, ⴱ, ⴳ, ⴳⵯ, ⴷ, ⴹ, ⴻ, ⴼ, ⴽ, ⴽⵯ, ⵀ, ⵃ, ⵄ, ⵅ, ⵇ, ⵉ, ⵊ, ⵍ, ⵎ, ⵏ, ⵓ, ⵔ, ⵕ, ⵖ, ⵙ, ⵚ, ⵛ, ⵜ, ⵟ, ⵡ, ⵢ, ⵣ, ⵥ) comme alphabet de la langue amazighe (Ameur *et al*., 2004). Avant sa standardisation informatique, l'analyse des sites amazighs sur la toile a montré que malgré qu'il existe une abondante documentation sur l'amazighe, il y a très peu de sites qui l'utilisent comme langue de communication. De plus, sa présence sur la toile en caractères tifinaghes semblait occasionnelle. La modélisation informatique de l'amazighe faisait appel à d'autres alphabets, en particulier l'alphabet arabe et l'alphabet latin enrichi par des caractères spéciaux empruntés de l'Alphabet Phonétique International (API). Par contre, les transcriptions en tifinaghe étaient généralement insérées sous forme d'images. Afin de surmonter ce problème, il a fallu codifier le système alphabétique tifinaghe, afin de l'intégrer dans tous les supports de communication. Ainsi, dans l'attente de la reconnaissance officielle de l'alphabet tifinaghe comme partie du plan multilingue de base, par l'ISO, (International organization for standardization) en 2004, et de l'amendement de l'Unicode, en 2005, qui est la solution générique de la standardisation des caractères tifinaghes (Andries, 2008), une solution rapide basée sur des normes de codage de caractères existantes, a été entreprise. Cette solution était basée sur le codage ASCII étendu ANSI (assurant l'encodage des caractères latins) et sur un ensemble de polices de caractères dédiées à la graphie tifinaghe.

2.2. Le Tifinaghe dans les technologies mobiles

Avant de réfléchir sur l'implantation d'une nouvelle application qui respecte un certain nombre de critères, il s'avère nécessaire de commencer par un diagnostic de l'existant, qui met le point sur les caractéristiques des applications existantes sur le marché, supportant le tifinaghe. Ce qui donne l'opportunité de réaliser un ensemble de comparaisons entre celles-ci et l'application à implanter.

En recherchant le terme tifinaghe dans l'Android Market de google[1], nous trouvons des applications sous différents types : des dictionnaires, des applications d'apprentissage de l'alphabet, ainsi que des applications de divertissement. Mais, nous restreignons notre analyse sur les dictionnaires mobile afin de donner plus de concrétisation à notre étude.

Mot recherché	Dictionnaires
Tifinaghe	Amawal Dictionnaire - Madghis U'Madi Dico Tifi/Ar/Fr : ASoftKorp
Amazighe (Berbère/berber)	Amawal - Dictionnaire Berbère : Abdessamad Idrissi

Tableau 1 : Liste des dictionnaires mobile existants en octobre 2013

D'après l'analyse des applications citées dans le Tableau 1, nous pouvons constater que la majorité des applications existantes ne disposent pas d'une zone de recherche avec la possibilité d'aide à la recherche via un module d'auto-complétion. Elles ne permettent pas une recherche avancée assurant la spécialisation des recherches dans un domaine déterminé. En outre, elles n'offrent pas la description phonétique des mots amazighes.

En se basant sur ces éléments, nous avons élaboré une application d'un lexique amazighe pour Android « LEXAM » écrit en tifinaghe, qui a pour fonctionnalités : la recherche et la consultation des mots, le partage dans les réseaux sociaux et l'envoi d'une demande d'ajout de nouvelles entrées afin d'enrichir la base de données de l'application. L'application permet aussi de créer une liste des mots favoris et offre l'option d'aide à la recherche, en permettant l'affichage d'un ensemble de suggestions de mots proches de ceux entrés par l'utilisateur.

3. Le lexique amazighe pour mobile

3.1. Methodologie

Après l'analyse de l'existant, nous avons pris la décision d'implanter une application mobile pour le lexique amazighe, qui sert de moyen pour permettre à l'amazighe de suivre cette nouvelle tendance de la nouvelle technologie mobile. Ainsi, l'élaboration de cette application est passée par plusieurs phases, à savoir :

- La conception des interfaces graphiques pour les fonctionnalités suivantes : l'accueil, la recherche, la consultation, et la demande d'ajout une nouvelle entrée (Voir Figure 1).

1 https://play.google.com/store/search?q=tifinagh, consulté en mars 2013

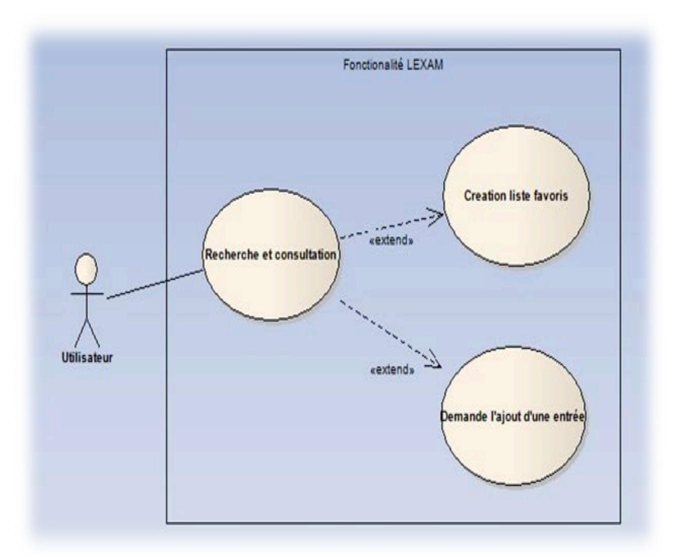


Figure 1 : Cas d'utilisation du LEXAM

- La création d'une base de données issue de celle du référentiel de la terminologie amazighe (Al Azrak et Elhamdaoui, 2011) dans le système HSQLDB (HyperSQL DataBase) [2], qui est un système de gestion de base de données relationnelle écrit en Java.
- Le développement de la couche service selon les étapes suivantes : Extraction des donnes via la couche modèle selon l'action de l'utilisateur dans la couche vue, sachant que nous avons adopté le patron Modèle-Vue-Contrôleur (MVC)[3].
- Contrôle de l'affichage des données dans la couche vue.
- Consommation du service de la translitération des données écrites en tifinaghe à des données écrites en caractères latins ou arabes selon l'action de l'utilisateur (Ataa Allah et Frain, 2013).

3.2. Développement

Vu que les versions Android actuelles n'intègrent pas la police tifinaghe Unicode, il s'avère nécessaire, avant d'entamer le développement du dictionnaire, de créer un clavier tifinaghe assurant à l'utilisateur la possibilité de rechercher un terme écrit en tifinaghe.

2 http://en.wikipedia.org/wiki/HSQLDB
3 http://fr.wikipedia.org/wiki/Mod%C3%A8le-vue-contr%C3%B4leur

3.2 .1. Clavier

L'écriture en tifinaghe est une étape importante, qui simplifiera et accélérera la consultation du contenu du dictionnaire. Néanmoins, afin de surmonter le problème de la non intégration de la police tifinaghe Unicode dans le noyau des versions actuelles d'Android, nous avons procédé à l'élaboration d'un clavier virtuel, via un ensemble d'étapes, dont nous citons :

- La création d'un fichier XML englobant la structure du clavier, dont l'aperçu est ci-dessous:

```
<?xml version="1.0" encoding="utf-8"?>
<Keyboard xmlns:android="http://schemas.android.com/apk/res/android"
    ……">
        <Row>
                <Key android:codes="11568" android:keyIcon="@drawable/a"/>
                <Key android:codes="11569" android:keyIcon="@drawable/b"/>
            ………
        </Row>
      <Row>
            ….
      </Row>
</Keyboard>
```

- La création d'un package ma.ircam.clavier dans le répertoire src de la solution et y mettre une classe KeyboardCustom en surchargent la méthode suivante :

```
@Override
        public void onKey(int primaryCode, int[] keyCodes)
                {
                }
```

- La création, via le code suivant, d'une vue du clavier (KeyboardView) dans l'activité[4] qui sera utilisée pour afficher le clavier :

```
  <android.inputmethodservice.KeyboardView
android:id="@+id/keyboardview"
android:layout_width="fill_parent"
android:layout_height="wrap_content"
android:focusable="false"
android:focusableInTouchMode="true"
android:visibility="gone" />
```

4 Une activité d'un système mobile est l'équivalent d'une fenêtre ou d'une page Web.

- L'instanciation de la class Kustomkeyboard dans le contexte associé à l'activité chargée d'afficher le clavier, par l'instruction suivante :

```
CustomKeyboard mCustomKeyboard= new CustomKeyboard(this,R.id.keyboardview,
R.xml.hexkbd );
```

Et l'association du clavier au texte voulu dans l'activité principale, par l'instruction suivante :

```
mCustomKeyboard.registerEditText(R.id.autoCompletemotarab);
```

3.2 .2. Affichage du Tifinaghe

Les composants graphiques offerts par le SDK (Kit de développement, Software Development Kit) d'Android sont classés selon les catégories suivantes : Form widget (bloc de formulaire), Text Field (zone de texte), Layout (zone de composants), Composite, Images et Média. La méthode appliquée pour afficher le caractère tifinaghe varie selon la catégorie du composant graphique :

- Pour les Form widgets et les Text Fields, il faut instancier la class Typeface et récupérer l'identifiant (id) du composant dans lequel sera affiché tifinaghe, en utilisant le code suivant :

```
Typeface fontTIF =
Typeface.createFromAsset(this.getAssets(),"fonts/T_I_UNICODE.ttf");
TextView titre=(TextView)this.findViewById(R.id.txt_titreAM);
titre.setTypeface(fontTIF);
```

- Pour le composite, il existe des listes simples, où chaque occurrence est constituée d'un seul champ, proposées par défaut sous Android. Ces listes permettent d'afficher les occurrences sous la forme d'un texte. Or dans le cas où les listes sont composées de plusieurs champs, l'affichage nécessite un traitement personnalisé. A titre explicatif, nous présentons l'exemple ci-dessous pour illustrer comment mettre en place une vue un peu plus complexe contenant, entre autres, une zone de texte dédiée au texte de la langue amazighe écrit avec la police tifinaghe.

Les éléments de la liste sont affichés dans un layout du type ListView (liste de vues). Dans une première étape, nous commençons par l'insertion d'un layout dans l'activité principale. Puis, nous constituons les occurrences de cette liste selon les étapes suivantes :

- Définir un modèle, illustré par l'exemple ci-dessous, qui organise et détermine les champs des occurrences de la liste.

```
<?xml version="1.0" encoding="utf-8"?>
    <LinearLayout xmlns:android="http://schemas.android.com/apk/res/android"
          android:id="@+id/LinearLayout2" ...>

  <TableLayout
          android:layout_width="match_parent"
          android:layout_height="match_parent">
          <TableRow
                    android:id="@+id/tableRow2">
                    <TextView
                              android:id="@+id/textamazighe"
                              android:layout_width="0dp"
                              android:layout_height="wrap_content"
                              android:text= "ⴰⵙⵓⵖⵍ" />
                    ............

           </TableRow>
          .........................
    </TableLayout>
```

- Définir un adaptateur (Adapter) pour gérer l'affichage de la ListView. Généralement, le SDK d'Android propose des classes de type Adapter, qui permettent d'afficher les données sur les vues selon un modèle déterminé. Ainsi, en héritant de la classe liée à la ListView, nous créions une nouvelle classe 'TermAdapterAm' décrite ci-dessous :

```
public class TermAdapterAm extends ArrayAdapter<Terme> {
        Terme t;
        public TermAdapterAm(Context context, int resource) {
                super(context, resource);
        }
        @Override
        public View getView(int position, View convertView, ViewGroup
        parent) {
                View resultat=convertView;
                if(convertView==null)
                {
                        resultat=LayoutInflater.from(getContext()).inflat
                e(R.layout.modele_terme_zgh, parent,false);
                }

                TextView txttifinaghe = (TextView) resultat.findViewById
                (R.id.textamazighe);
                 // set the font " TIFINAGHE " in a textView
                Typeface.createFromAsset(getContext().getAssets(),"fonts/
                T_I_UNICODE.ttf");

                txttifinaghe.setTypeface(font);
                        ………..

                //set an existed font in a textView
                        Typeface fontlatin = Typeface.createFromAsset
                (getContext().getAssets(),"fonts/Cardo-Italic.ttf");
                txt4.setTypeface(fontlatin);

                //set the result in an Array adapter
                t=getItem(position);
                txt1.setText(t.getMotAmazighe());
                txt2.setText(t.getMotFrancais());
                txt3.setText("  "+t.getMotArabe());
                txt4.setText("/ "+t.getDetail()+" /");

                return resultat;

        }
```

- Instancier la classe 'TermAdapterAm' au sein de la méthode surchargée 'onCreate' de l'activité principale, selon le code suivant :

```
public class ActivityZgh extends Activity implements OnItemSelectedListener
        {
        @Override
                protected void onCreate(Bundle savedInstanceState)
                {
                        super.onCreate(savedInstanceState);
                        //initialization of a listView adapter
                        termAdapter=new TermAdapterAm(this,
                        R.layout.modele_terme_zgh)
                }
```

3.3. Démonstration du LEXAM

Ecran	Description
	Le lexique amazighe marocain «LEXAM» couvre des domaines de la vie courante et moderne : médias, administration, art, environnement, civilité, droit et justice, éducation, et bien d'autres. La version actuelle est trilingue (Amazighe, Arabe, Français) . Elle contient plus de 4000 entrées.
	L'application contient un clavier virtuel amazighe, assurant la recherche en caractère tifinaghe avec l'option de l'auto-complétion.
	Le résultat de recherche permet l'accès a toutes les entrées qui commencent par la lettre saisie par l'utilisateur, selon la structure ci-dessous : Mot en tifinaghe /*Transcription latine* / Traduction en français Traduction en arabe

4. Conclusion

Cet article décrit l'élaboration d'une application mobile pour un dictionnaire amazighe. Cette application regroupe un ensemble d'améliorations par rapport aux dictionnaires existants, afin de simplifier son exploitation et son utilisation par différents utilisateurs.

Dans la perspective d'améliorer ce travail et toutes les applications mobile dédiées à la langue amazighe, en général, nous envisageons de créer un clavier tifinaghe intégré dans les appareils mobile et d'ajouter la police tifinaghe IRCAM aux prochaines versions du noyau Android. Ce qui facilitera, certainement, l'élaboration des applications supportant le tifinaghe aux développeurs et le manipulation du clavier tifinaghe aux utilisateurs.

Références

Ameur M., Bouhjar A., Boukhris F., Boukous A., Boumalk A., Elmedlaoui M., Iazzi E. et Souifi H. (2004). Initiation à la langue amazighe, Maroc : IRCAM.

Andries P. (2008), Unicode 5.0 en pratique : Codage des caractères et internationalisation des logiciels et des documents, France : Dunod.

Ataa Allah F. et Frain J. (2013). Amazigh Converter based on WordprocessingML. *6ème conférence en Language & Technology*, 7-9 décembre 2013, Poznań, Pologne.

El Azrak N. et Elhamdaoui A. (2011). Référentiel de la terminologie amazighe: outil d'aide à l'aménagement linguistique. *4ème atelier international sur l'Amazighe et les NTICs*, 24-25 février 2011, Rabat, Maroc.

Hachid M. (2001). Les premiers berbères, Entre Méditérranée, Tassili et Nili, Aix-en-Provence-Alger : Edisud-Ina-Yas, pp. 173-190.

Lhote H. (1979). Les gravures de l'Oued Mammanet. Sénégal : Les Nouvelles Editions Africaines.

Skounti A., Lemjidi A. et Nami E. M. (2003). Tirra aux origines de l'écriture au Maroc, Maroc : IRCAM.

Recherche en texte intégral

Anass Smaili Ali Lasfar Mohamed Sbihi
LASTIMI
Ecole Supérieure de Technologie de salé
Université Mohammed V
{anass.smaili, mohamed.sbihi}@yahoo.fr
ali.lasfar@gmail.com

Résumé

Depuis les années 90, l'outil informatique a continué à prendre une place privilégiée et indispensable dans l'élaboration du travail administratif. Il est incontestable que l'exploitation des documents numérisés est beaucoup plus simple en termes de recherche.

L'automatisation, la disponibilité, la rationalisation et la rapidité de la réponse aux requêtes demandées par les utilisateurs, sont des attentes primordiales dans chaque base de données. Ces critères principaux nous amène à déterminer le degré et la puissance d'un moteur de recherche. Cet article présente une comparaison entre deux moteurs de recherche afin d'aboutir à avoir une réponse dans un temps d'attente court.

1. Introduction

L'introduction de la langue amazighe dans le système éducatif marocain demande une compétence juridique et technique dans la matière à légiférer. Cela dévoile, une mosaïque comprenant des connaissances extrêmement variées : constitution, droit, jurisprudence, conventions, etc. Cependant, l'outil informatique a facilité l'exploitation de ces fonds colossaux, qui sont maintenant organisés dans des documents numérisés couramment consultables.

L'existence d'un service de recherche ne fait aujourd'hui plus de débat. En revanche, la définition de ce que ce que cette fonctionnalité peut offrir, la détermination de sa nature ne cessent d'interroger.

Il est fréquent que des besoins, considérés comme identiques, dérivent des services différents. On se pose alors la question de l'utilisation de la fonction de recherche classique du langage SQL « like », la solution adéquate pour des bases de données de petite taille, cependant, la fonction SQL « like » est « irréalisable » pour des bases de données de grande taille.

D'autres mécanismes sont mis en effet pour réponde aux bases de données volumineuses.

La difficulté à avoir des résultats, après les traiter, par la suite en connaître la nature, est inappréciable par l'utilisateur, bien que le moteur de recherche contienne une masse importante d'information. Ainsi, l'objectif de cet article est de mettre le point sur le rôle du critère du temps de réponse aux requêtes utilisateurs, en dressant une comparaison entre deux moteurs de recherche.

2. Architecture matérielle/logicielle utilisée

Nous proposons dans notre plateforme de test, deux serveurs. Le premier qui abrite l'application développée qui interroge la base de données, le second qui contient la base de données. Les deux serveurs contiennent deux processeurs et 24 Go de mémoire. Nous supposons que le système de gestion de base de données est SQL Server 2008.

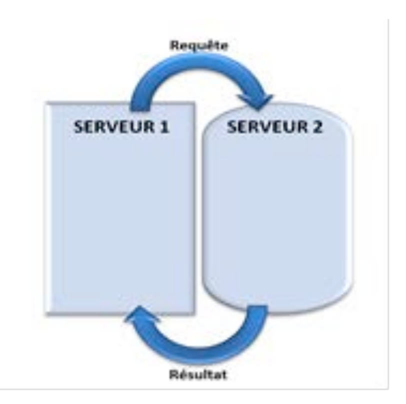


Figure 1 : Communication entre le serveur d'application et le serveur de base de données

Nous supposons aussi que le schéma de la base de données contient une seule table composée de trois champs, le premier contient une clé primaire de type entier, le deuxième champ contient des lignes de textes de type caractère (plus de 255 caractères) et le troisième champ contient un numéro de type caractère (moins de 255 caractère).

3. Présentation des outils

3.1. Lucene

Lucene est un moteur de recherche (Cai *et al*., 2006 ; Yin, 2009 ; Zhou et Wang, 2009) développé en java destiné pour la recherche du texte intégral. Il se divise en deux briques principales : indexation et recherche.

La brique réservée à l'indexation est composée de plusieurs classes (Yaofang, 1992 ; Cai *et al*., 2006) qui se chargent de créer des entrées des différents termes trouvés dans la source de données. Un passage par un analyseur qui extrait les racines des mots rend le contenu de l'index plus optimal et permet le bon calcul du nombre d'occurrence du terme.

La brique destinée à la recherche (Yaofang, 1992) au sein de lucene comprend des classes qui se chargent d'explorer le fichier « Index », créé auparavant suite au processus d'indexation (Su et Widom, 2005). L'opération consiste à traiter la chaîne de caractère recherchée selon sa décomposition analytique, puis classer les résultats en se basant sur un conteneur dédié à cette tâche.

Le cas d'utilisation de lucene a demandé de référencer d'abord, les librairies relatives à lucene.net, via le package manager console qui doit être intégré à Visual Studio.

La première étape consiste à la création/liaison de la source de données (base de données SQL server dans notre cas), afin de pouvoir créer la classe d'exploitation du moteur lucene, cette base de données contient des fonctions permettant les fonctionnalités suivantes:

- La recherche
- L'indexation
- La consultation de l'index
- L'analyseur des requêtes lucene
- Le mappage index lucene avec les enregistrements des tables au niveau de notre BD
- Mise à jour de l'index
- Suppression de l'index ou d'un élément de l'index
- Optimisation de l'index
- L'ajout d'un document à l'index

Il ne faut pas omettre les attributs de la classe, notamment le chemin du répertoire qui est censé contenir l'index et le chemin du fichier contenant la liste des mots vides.

Le terme de la recherche est une chaîne de caractère qui suit tout d'abord un traitement, par la suite il est assigné pour supprimer les espaces au début et à la fin, ensuite cette chaîne est dupliquée. Dans la première : les espaces utiles sont remplacé par le mot clé 'AND' par contre dans la deuxième ils sont remplacés par le mot clé 'OR' ; ensuite une requête est formulée (à partir de ces deux chaînes de caractères).

3.2. FTI Sqlserver 2008 R2

SQL server contient un moteur d'indexation et de recherche en texte intégral (Su et Widom, 2005 ; Fang *et al*., 2008) composé de deux processus : « SQL server » et « hôte de démon de filtre ».

Le processus SQL server (Hamilton et Nayak, 2001) contient à la fois le processeur de SQL Server de requête, qui compile et exécute les requêtes SQL, et le moteur de texte intégral, qui compile et exécute les requêtes du texte intégral. Par contre, l'hôte de démon du filtre est un processus externe à SQL Server démarré par le moteur d'indexation et de recherche en texte intégral. Il exécute les composants de recherche en texte intégral, chargés d'accéder aux données des tables, de les filtrer et d'en effectuer l'analyse lexicale, ainsi qu'à identifier la racine des mots de l'entrée de la requête (Wang et Zhang, 2005).

En bénéficiant de ce qu'offre le moteur de recherche propre à SQL server, nous avons commencé tout d'abord par lier le projet à la base de données SQL server afin d'interroger les données.

Framework EF (entity framework 4.0) a été utilisé pour générer le modèle qui constitue la base de données du projet. Ensuite un contrôleur de recherche a été créé pour gérer les événements et la synchronisation des vues.

Les principales fonctions qui ont été développées concernent le bouton « rechercher » :

L'utilisateur doit saisir un terme ou une phrase qui fait l'objet de sa recherche ensuite en appuyant sur le bouton rechercher la fonction 'index()' du contrôleur 'searchFr' est appelée. Le terme (ou la phrase) est traité, dans la fonction 'index()', pour savoir si le terme correspond à un numéro de texte ; Si c'est le cas, la recherche va porter sur le champ de la base de données qui contient les numéros de texte en appelant la procédure appropriée, sinon la recherche se fait sur le champ qui contient les lignes du texte.

Dans les deux cas, la chaine de caractère recherchée passe par un bloc de traitement qui enlève les espaces au début et à la fin de la chaîne et remplace les espaces par les mots clé 'AND' ou 'OR'. Deux paramètres sont passés à cette fonction pour faciliter la pagination :

Paramètre 1 : qui sert à sauter les enregistrements déjà consultés.

Paramètre 2 : qui prend les enregistrements de la page sélectionnée.

A la fin de la recherche, la liste est retournée à la vue pour affichage des résultats.

3.3. Echantillon

La base de données que nous avons prise contient 300 000 enregistrements.

Le tableau (Tab1) ci-dessous comporte trois colonnes qui comprennent les temps de réponse respectivement d'un moteur de recherche basique qui utilise la fonction « like », moteur de recherche avec lucene et moteur de recherche avec FTI SQL server.

Chaque ligne contient la requête passée sur les trois moteurs de recherche.

Requête	Basique	Lucene	FTI SQL server
Sécurité	2.53	0.024	0.042
Sécurité nationale	3.94	0.140	0.098
Constitution sécurité nationale	8.80	0.151	0.150
décret loi	5.1	0.206	0.272
Décret loi dahir arrêté décision	13.55	1.082	1.386

Tableau 1: Temps de réponse entre les interrogations des trois moteurs de recherche

3.4. Résultat

Le contexte dans lequel s'inscrit ce projet est le rapprochement de l'information, liée aux différents domaines de recherche dans la langue Amazigh, au citoyen. L'autre élément contextuel tient à l'obtention du résultat en un temps infime.

Selon le diagramme ci-dessous (fig.2), le temps de réponse du moteur de recherche « basique » est largement plus important que ceux de lucene et fti SQL server. Cette différence est due en grande partie à l'absence d'une indexation, qui s'accompagne par le manque d'utilisation de la recherche en texte intégral.

L'analyse du résultat de lucene et fti SQL server montre l'importance concomitante du délai court du traitement et de la réponse des requêtes.

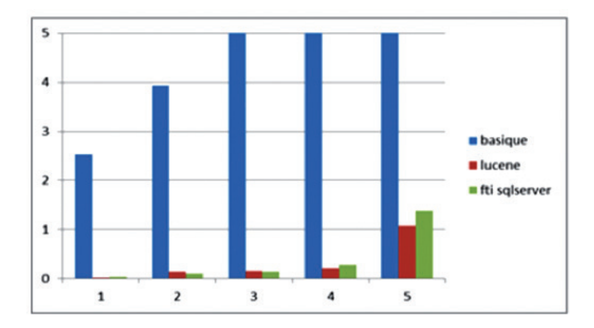


Figure 2 : Temps de réponse en interrogeant les trois moteurs de recherche (X= temps en seconde ; Y : numéro d'échantillon)

4. Conclusion et perspectives

L'écart du délai de réponse entre le moteur de recherche « basique » et lucene ou fti SQL server est très important. Dès lors, le passage à la recherche en texte intégral en utilisant l'un de ces outils est clair : indexer les données augmente la vitesse de traitement de la requête.

Malgré cela, cet argument ne saurait répondre à lui seul à la problématique de dégradation de temps de réponse. Cette dégradation s'observe dans tous les paramètres fonctionnels et techniques, depuis le commencement de la conception de ce projet.

Notre but dans les prochains articles est de présenter une conception générale de la base de données (Smaili, 2012) ; de bénéficier des différents paramètres offerts par ces deux outils pour améliorer la performance de notre programme en terme de rapidité de temps de réponse; d'augmenter la pertinence de la recherche en incluant la sémantique (Sauermann *et al*., 2005; Hogan *et al*., 2007), et enfin mettre en place une plateforme de test plus puissante et bien configurée.

Références

Cai J., Guo Y., Wang L. (2006). Design and Implementation of School Search Engine Based on Lucene.Net. *Computer Technology and Development*.

Fang Y., Friedman M., Nair G., Rys M., Schmid A.E. (2008). Spatial indexing in microsoft SQL server. *SIGMOD*, pp. 1207-1216.

Hamilton J. R., Nayak, T. K., (2001). Microsoft SQL server full-text search. *IEEE Data Engineering Bulletin*. 24(4): 7-10.

Hogan A., Harth A., Umbrich J., Decker S. (2007). Towards a scalable search and query engine for the web. *International Conference on World Wide Web*. pp. 1301-1302.

Sauermann, L., Bernardi, A., Dengel, A. (2005). Overview and outlook on the semantic desktop. *Semantic Desktop Workshop at the International Semantic Web Conference*, Galway, Ireland, November 6th 2005.

Smaili A., Lasfar A., Sbihi M. (2012). A Word Image Matching: Case of the Official Bulletin of Kingdom of Morocco. *International Journal of Computational Linguistics Research*. pp. 143-149.

Su Q., Widom J. (2005). Indexing relational database content offline for efficient keyword-based search. IDEAS.

Wang S., Zhang K-L. (2005). Searching databases with keywords. *The Journal of Computer Science and Technology*. 20(1): 55–62.

Yaofang G. (1992). On The Full-Text Retrieval System. *New Technology of Library and Information Service*.

Yin W. (2009). Study on Lucene-Based Application. *Journal of Taizhou Polytechnic Institute*.

Zhou J., Wang D. (2009). Research and Application of Full-Text Retrieval Search Engine Based on Lucene. *Journal of the Qiannan Normal College for Nationalities*.

Présentation d'un Atlas linguistique parlant

Noura Tigziri
Laboratoire Aménagement et enseignement de la langue amazighe,
Université Mouloud Mammeri de Tizi-Ouzou.
Nora.tigziri@gmail.com

Résumé

Notre projet consiste en la mise en place d'une banque de données de corpus oraux, numérisés, transcrits et annotés pour la langue amazighe qui soit exploitable à des fins scientifiques s'adressant principalement aux enseignants chercheurs linguistes. Nous souhaitons récolter un corpus suffisamment large pour qu'il soit représentatif de la langue, et afin qu'il permette sa sauvegarde sous forme de ressource linguistique.

Le premier objectif est de mettre à disposition de linguistes une ressource linguistique ce qui implique des conséquences sur la manière de définir les métadonnées et les annotations. Cette recherche est aussi une occasion de documenter le kabyle parlé dans toutes ses variétés, sous toutes ses formes géographiques. Son intérêt réside aussi du fait que cette ressource linguistique sera accessible via le web, et conduire en première application à la réalisation d'Atlas linguistique parlant dont nous présentons une ébauche dans cette présentation.

Les choix technologiques ont été exposés lors de la communication présentée au colloque NTIC de 2011 de l'IRCAM.

Au bout de trois années de recherche et d'enquêtes, nous avons réalisé :

100 points d'enquête, et 100 enregistrements de 20 mn chacun pour la plupart transcrits. Des champs lexicaux ont été aussi pris en charge par nos enquêtes.

La première application que nous envisageons est l'élaboration d'un atlas linguistique parlant. C'est-à-dire dans un premier temps représenter nos points d'enquête sur une carte et accompagner ces points d'enquête d'un enregistrement sonore, de la transcription du son et plus tard en étudier la représentation acoustique pour tout ce qui concerne les phénomènes prosodiques du kabyle.

1. Géolocalisation sur QGis des corpus recueillis.

Après avoir tracé les outils théoriques et méthodologiques pour notre travail, nous avons tenté un essai de géolocalisation des points d'enquêtes sur Qgis avec un certain nombre de paramètres à prendre en charge.

La première phase consiste à télécharger et installer la dernière version de QGis ('2.2 Valmiera' février 2014): http://www.qgis.org/fr/site/

Pour ce faire, nous avons suivi les étapes suivantes :

L'utilisation d'un logiciel de cartographie numérique tel QGis implique souvent l'utilisation de coordonnées géographiques, tel qu'une valeur de latitude ou de longitude.

Pour avoir ces valeurs de longitude et de latitude pour l'Algérie et plus précisément la Kabylie, domaine de nos enquêtes linguistiques, nous avons importé des fonds de cartes avec les coordonnées des points à partir de : http://www.diva-gis.org/gdata

En sélectionnant Algeria dans le champ Country, nous avons téléchargé les fichiers suivants:

Administrative areas:	DZA_adm.zip (frontières administratives)
Inland water:	DZA_wat.zip (lacs et cours d'eau)
Roads:	DZA_rds.zip (routes)
Elevation (country mask):	DZA_msk_alt.zip (altitudes)

Ces fichiers ont été ensuite intégrés dans QGis par l'intermédiaire d'options 'intégrer une couche vectorielle'. Cette opération a été répétée pour chaque couche dont on avait besoin et qu'on pouvait exploiter plus tard pour l'analyse et l'interprétation des résultats.

Ainsi à partir d'*Administrative areas*, nous avons pu disposer de tracés de l'Algérie en tant qu'Etat (Figure 1), des Wilayas (Figure 2) et des communes (Figure 3), subdivisions administratives en Algérie.

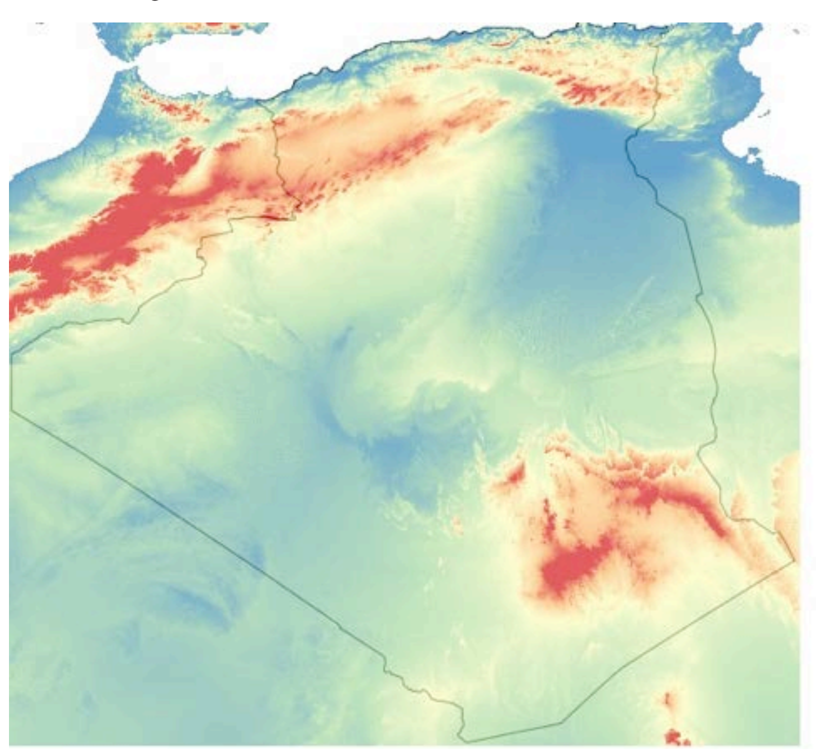


Figure 1 : ETAT

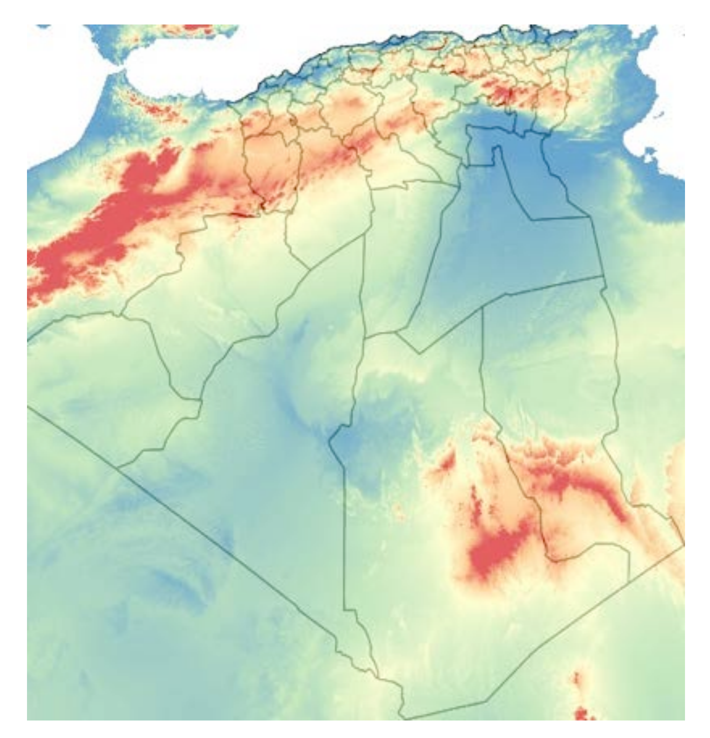


Figure 2 : Wilayas

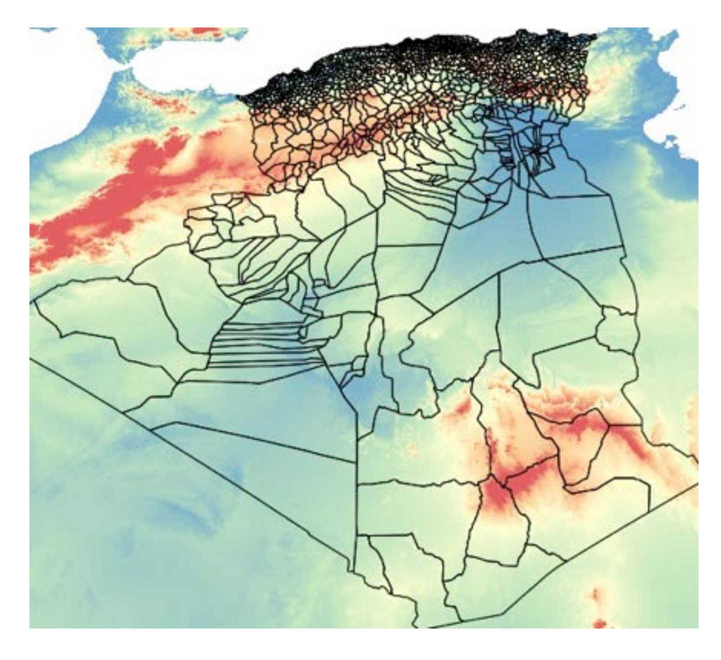


Figure 3 : Communes

Nous avons aussi intégré les plans et cours d'eau (Figure 4) et les routes (Figure 5) :

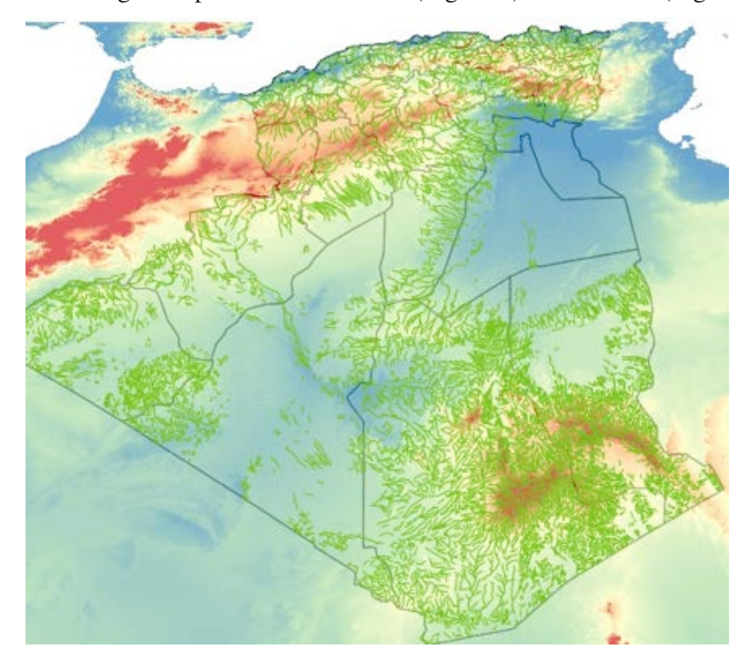


Figure 4 : Cours d'eau

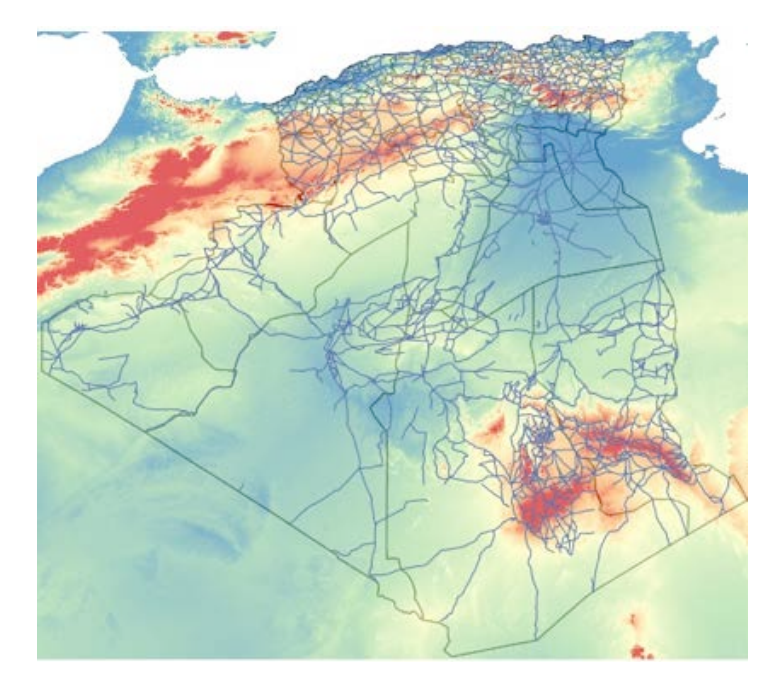


Figure 5 : Routes

Notre travail consistant à représenter des points d'enquête sur le territoire national. Il est impératif d'ajouter des noms de lieux sur ce fond de carte. Pour cela nous retournons sur le site : http://www.diva-gis.org/gdata

Et toujours, en sélectionnant *Algeria*, nous téléchargeons cette fois-ci le fichier DZA_gaz.zip dans subject en sélectionnant '*Gazetteer*', dont le fichier extrait DZA.dbf représente un tableau ou base de données au format dBase IV. La localisation d'un point sur ces cartes repose sur les données de la latitude et de la longitude, relativement au système géodésique de référence WGS84 (World Geodetic System 1984), probablement le plus utilisé (en particulier par les GPS et par Google Earth).

Ce fichier comporte des doublons et parfois le même lieu écrit sous plusieurs écritures, aussi avant de l'utiliser, il faut le trier pour ne garder que les données qui nous intéressent.

Nos premières enquêtes portent principalement sur la Kabylie, aussi on trie encore de manière à n'avoir pour une première phase que les localités (Figure 6) concernant cette région. Et on obtient :

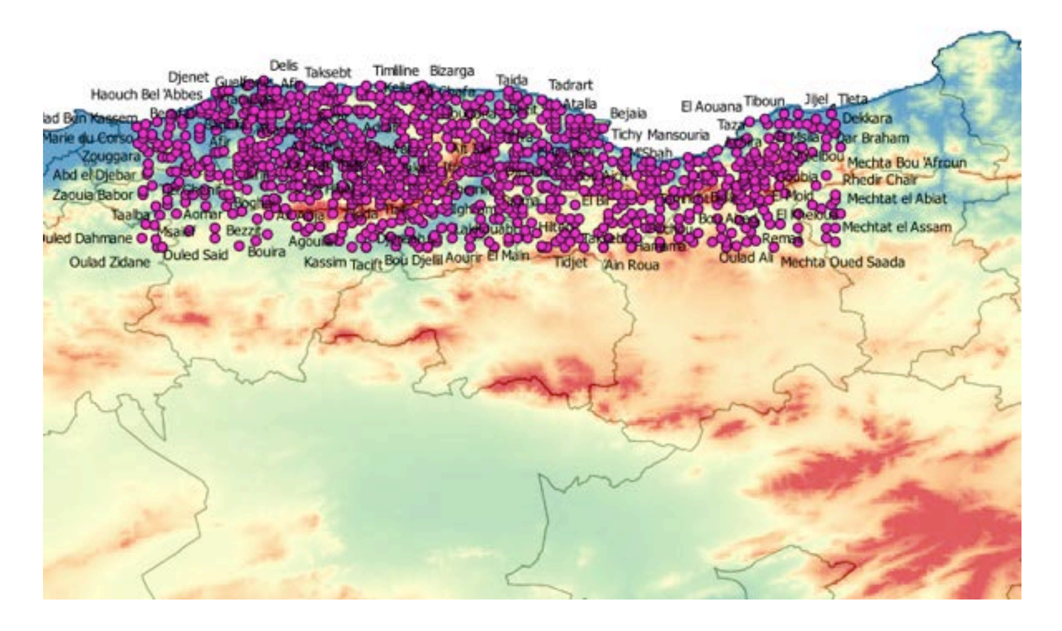


Figure 6 : Localités

2. Représentation de quelques points d'enquête :

Le tableau ci-après donne quelques points d'enquête avec les coordonnées de géolocalisation, à savoir la latitude et la longitude.

Latitude	Longitude	Name
36,819	4,213	Abizar
36,8235	4,292	Adrar n Takdhiâa
36,811	4,172	Afir
36,614	4,042	Aglagal
36,809	4,322	Aghrib
36,71	3,768	Ain-Sakoura
36,547	3,894	Ain Zaouia
36,579	4,041	Ait Abdelmoumen
36,556	4,273	Ait Ailem
36,57	4,451	Ait Ali oumohand
36,622	4,1	Ait Bou-Yahia
36,58	4,329	Ait Hichem
36,629	4,116	Ait Khelfoune
36,711	3,584	Ait Welhadj
36,658	4,437	Ait-Zellal
36,616	4,76	Akabiou
36,804	4,323	Aghrib
36,619	4,042	Aglagal
36,6	4,004	Aagueni Bouffal
36,8	4,266	Agueni ouzaraz
36,5242	4,431	At Aziz
36,62	4,1	Ath Bou Yahia
36,673	4,413	Ait-Ghobri
36,683	3,85	Ath Hwelhadj
36,5268	4,277	Ath Sâada
36,59	4,307	Boudafal
36,791	4,229	Boudrar
36,8	4,266	Bouguedama
36,25	4,165	Bouira
36,5867	3,918	Boumahni
36,483	4,983	Bouzekout
36,85	4,123	El Kelaa
36,812	4,244	Ibdhach
36,604	4,07	Icardiwen
36,801	4,073	Ichikar
36,44	3,919	Ighechamen
36,745	3,9	Igherbiyen

36,626	3,892	Ighil El-Vir
36,5685	4,261	ighil bougni
36,553	4,043	Ighil Imoula
36,627	3,851	Ighil-El-Vir
36,668	4,52	Iguersafene
36,633	3,819	Imella?en
36,718	4,24	Imzizou
36,807	4,165	Isserradjen
36,6641	3,895	Ivouhrene
36,652	3,946	Izarouden
36,6163	3,912	Kantija
36,791	4,244	Laazib
36,734	4,045	Redjaouna
36,5979	4,458	Sahel
36,621	3,848	tafoughalt
36,549	4,171	tahechat
36,584	3,957	tajnant
36,613	4,331	Taka Ait Yahia
36,767	4,534	Talbent
36,8	4,35	tamassit
36,631	3,804	Tamdhiqt
36,89	4,109	tasalast
36,571	4,262	Tasga Melloul
36,7993	4,118	Tissegouine

Dans ce fichier excel, nous avons prévu trois autres données : doc, son, web.

La colonne doc nous permettra d'intégrer dans QGis un document « texte », la colonne son d'intégrer un fichier son, et celle relative à web permet d'aller chercher quel que soit le document à partir du web.

Remarques

La géolocalisation des points d'enquête reste un des points les plus fastidieux de notre travail. Ne disposant pas de GPS pour relever les coordonnées des points d'enquête, nous avons été obligé d'élaborer un certain nombre de stratégies dont les plus importantes sont les suivantes :

- Passer par la base de données de Google Earth ;
- Passer par d'autres bases de données comme celle de la National Geospatial-Intelligence Agency (NGA, cf. http://earth-info.nga.mil/gns/html/namefiles.htm);
- Chercher sur web le nom du village concerné et retrouver ses coordonnées, parfois données par des particuliers.

L'écriture des toponymes n'étant pas stable et la même à travers le temps et l'espace, nous avons rencontré énormément de difficultés à accomplir cette tâche. Ainsi par exemple, nous avons pour un même toponyme les écritures suivantes :

Ait Iraten, At Iraten, Ait-Iraten, At-Iraten, At Irathen, Ait Irathen, At-Irathen…

La carte obtenue de nos points d'enquête est donnée ci-après (Figure 7) :

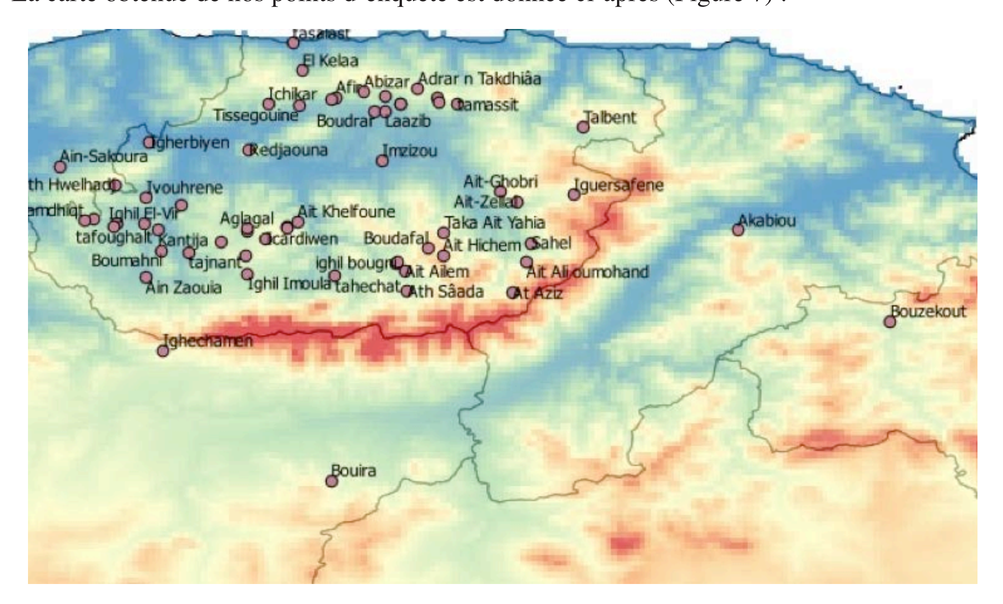


Figure 7 : Représentation spatiale des points d'enquête

L'association d'un son de sa transcription est lancée par une fonction de QGis appelée «Actions» résumée dans ce tableau (Figure 8) :

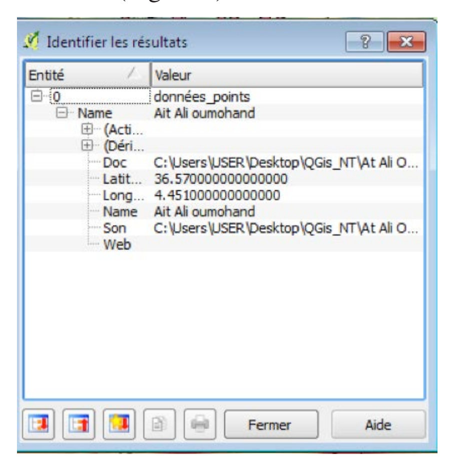


Figure 8 : Liste des actions activées dans QGIS

Ensuite, il suffit de choisir l'action qu'on veut lancer (son ou document) pour écouter l'enregistrement effectué ainsi que lire sa transcription sous forme de document word.

3. Conclusion

Ce projet nous tient à cœur pour plusieurs raisons. En premier lieu, nous disposerons d'une base de données de corpus oraux des différentes régions de la Kabylie et plus tard de toutes les régions berbérophones de l'Algérie.

Ces corpus peuvent être exploités dans plusieurs applications dont celle qui nous préoccupe le plus, l'aménagement.

En second lieu, chaque corpus sera accompagné « d'une pièce d'identité » ou fiche de collecte ainsi que de la transcription, ce qui constitue une ressource linguistique conséquente.

Bibliographie

Bachy S., Hambye P., Simon A. C. (2006). Tutoriel Praat, Cahier 2, Transcription et alignement du texte sur le son. Document non publié.

Bally C. (1935). Le Langage et la Vie, Zurich, Max Niehans (2ème éd.).

Blanche-Benveniste C. (2000). Transcription de l'oral et morphologie », Romania Una et diversa, Philologische Studienfür Theodor Berchem (Gille M. et Kiesler R. Eds). Tübingen : Gunter Narr, pp. 61-74.

Blanche-Benveniste C., Rouget C., Sabio F. (2002). Choix de textes de français parlé. 36 extraits. Paris, Honoré Champion.

Boersma P., Weenink D. (2007). Praat : doing phonetics by computer (Computer program, Version 4.5.16). Retrieved February 18, 2007, from http://www.praat.org.

Candea M. (2000). Contribution à l'étude des pauses silencieuses et des phénomènes dits «d'hésitation » en français oral spontané, Thèse non publiée.

Cheepen C. (1995). Discourse considerations in transcription and analysis. Leech G., Myers G., Thomas J. (Éds.), Spoken English on Computer. Transcription, Mark-up and Application, New York, Longman, pp. 135-143.

Dister A. (2007). De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL, Thèse de doctorat non publiée, Université de Louvain.

Dister A., Francard M., Geron G., Giroul V., Hambye P., Simon A. C., Wilmet R. (2006). Conventions de transcription régissant les corpus de la banque de données VALIBEL (http://valibel.fltr.ucl.ac.be, corpus oraux, conventions de transcription).

Duez D. (1997). La signification des pauses dans la production et la perception de la parole. *Revue Parole*, n°3-4, pp. 275-299.

Durand J., Laks B., Lyche C. (2002). "Protocole d'enquête", Bulletin PFC n°1 : Protocole, conventions et directions d'analyse, téléchargeable sur www.projet-pfc.net/, rubrique Accueil, Bibliothèque PFC.

Protocole, conventions et directions d'analyse. Toulouse, CNRS ERSS-UMR5610 et Université de Toulouse-Le Mirail, Bulletin PFC n° 1, 74 p.

Edwards J. A. (1995). Principles and alternative systems in the transcription, coding an mark-up of spoken discourse. Leech G., Myers G., Thomas J. (Éds.), Spoken English on computer. Transcription, Mark-up and Application, New York, Longman, pp. 19-34.

Ehmer O. (2006). The Tranformer – a tool for working with aligned language data. Communication présentée à la 3ème rencontre fribourgeoise de la linguistique sur corpus appliquée aux langues romanes, Freiburg-im-Breisgau, 14-17 septembre 2006.

Raingeard M., Lorscheider U. (1977). Édition d'un corpus de français parlé. Recherches sur le français parlé 1, pp. 14-29.

Simon A. C. (2001). Le rôle de la prosodie dans le repérage des unités textuelles minimales. Cahiers de linguistique française 23.

Simon A. C. (2004). La structuration prosodique du discours en français. Une approche multidimensionnelle et expérientielle, Berne, Peter Lang.

http://www.qgis.org/fr/site/

http://www.diva-gis.org/gdata

La transcription synchronisée texte/son des corpus oraux : Essai d'application sur un corpus kabyle

Ramdane Boukherrouf **Noura Tigziri**
Laboratoire Aménagement et enseignement de la langue amazighe
Département de langue et Culture Amazighes
Université Mouloud Mammeri de Tizi-Ouzou
Boukherouf@yahoo.fr Nora.tigziri@gmail.com

Résumé

Notre contribution consiste à présenter une petite tentative d'une application sur un corpus kabyle d'une transcription synchronisée des corpus oraux. En s'inscrivant dans le cadre du projet du Laboratoire des langues et Civilisations Orales[1], nous avons tenté d'annoter un échantillon d'un corpus kabyle, enregistré dans une situation concrète. Dans notre cas, nous avons opté d'annoter quatre niveau : texte intégral, traduction en français, paragraphes et énoncés.

1. Introduction

Après avoir constitué une banque de corpus oraux (enregistrements), nous voulons passer dans ce projet à leur exploitation. Pour cela, nous avons comme objectif de présenter ces corpus oraux alignés avec leur transcription.

Toute analyse linguistique de productions orales ne peut se faire qu'à partir d'une source sonore uniquement. En effet, pour qu'ils deviennent objet d'étude il est nécessaire de passer par leur mise par écrit. Aussi parler de traitement des corpus oraux, c'est prendre en compte de données primaires (les enregistrements) auxquelles il faut ajouter des données secondaires (leurs transcriptions) pour les rendre exploitables.

Mais travailler sur les corpus oraux et leurs transcriptions présupposent des choix notamment en ce qui concerne :

- La transcription à utiliser
- Le choix de noter ou pas la ponctuation

En effet, l'oral est très loin de l'écriture phonologique standardisée adoptée par les berbérisants. Les différents phénomènes observés dans la prononciation dans l'oral sont difficiles à systématiser comme il est ardu de distinguer des variantes de prononciation à la seule écoute du signal sonore ; de même que ce mode de transcription complique la tâche du transcripteur, surchargeant son travail par l'attention qu'il doit porter à ces phénomènes, dont la fréquence est souvent très élevée.

1 LACITO : http://lacito.vjf.cnrs.fr/

La ponctuation des corpus oraux :

Les corpus de langue parlée transcrits à des fins de recherche linguistique ne sont en général pas ponctués. On ne peut parler de correspondance stricte entre phénomènes prosodiques et ponctuation écrite. Il est difficile de savoir s'il faut mettre une virgule pour distinguer une petite pause, ou une grande pause…Même si on arrive à ponctuer certains énoncés plus faciles à segmenter comment s'en sortir avec des productions sonores dans lesquelles se multiplient les répétitions de mots, des inachèvements et autres phénomènes propres à l'oral ? D'ailleurs Blanche-Benveniste et Jeanjean plaident pour des transcriptions non ponctuées : «La ponctuation, si on la met trop tôt, préjuge de l'analyse syntaxique et impose un découpage sur lequel il est difficile de revenir.» (Blanche-Benveniste et Jeanjean, 1987 : 139).

2. Cadre théorique

Dans le projet nous allons utiliser le code source du LACITO[2] en langage xml. La segmentation du fichier son en sections sera faite à l'aide du logiciel «Audacity». Pour pouvoir exploiter ce logiciel, il est indispensable que les fichiers sons soient en format wav.

Les annotations : Annoter un document consiste à lui attacher un commentaire, une note ou une explication. Dans notre cas, il s'agit de transcriptions et traductions (à voir plus tard) du contenu de nos enregistrements ou corpus oraux.

Les annotations peuvent se présenter sous ces formes :

- Un document PDF, résultat de la numérisation de notes manuscrites ;
- Un document au format texte ;
- Un document structuré au format XML....

Les fichiers d'annotation tels qu'ils ont été définis dans le cadre de la Collection Pangloss possèdent 4 éléments de base :

- Une balise ***TEXT*** qui englobe toutes les informations ;
- Une balise ***S*** (sentence) qui permet de délimiter chaque "phrase" du texte ;
- Une balise ***W*** (Word) pour un découpage de la phrase en mot ;
- Une balise ***M*** (morphème) pour un découpage des mots en morphèmes.

Pour chaque élément, il est possible d'ajouter une traduction ou une glose grâce à la balise ***TRANSL*** (translation).

Une balise ***NOTE*** permet également d'ajouter des commentaires pour le texte ou pour chaque phrase. Enfin, puisque la base de toutes ces annotations est l'enregistrement audio, il est possible de synchroniser le fichier audio et le fichier d'annotations grâce à la balise ***AUDIO***.

3. Cadre pratique

Après avoir présenté ci-dessus la partie théorique de notre projet, nous tenterons dans ce point une application sur un échantillon d'un corpus kabyle annoté avec quelques propositions de balises.

2 http://lacito.vjf.cnrs.fr/

3.1. Présentation du corpus

Notre corpus a été enregistré dans une situation réelle. Il s'agit d'un meeting, d'une durée de quarante-cinq minutes, prononcé le deux (2) mai 2002 à quatorze heures et quinze minutes au stade Oukil Ramdane à Tizi-Ouzou par le président du parti de Rassemblement pour la Culture et la Démocratie, le docteur Saïd Sadi, lors des élections législatives de 2002.

Pour la transcription, nous avons adopté une notation à tendance phonétique dans le but de mettre en valeur les spécificités de l'oral, notamment les réalisations contextuelles ou combinatoires.

En outre, afin de reproduire plus au moins fidèle le texte oral, nous avons pris en considération les hésitations, les répétions et les différentes pauses qui segmentes les différentes unités textuelle du corpus.

Dans notre corpus nous avons noté trois types de pauses :

- Les pauses courtes [/] : ce sont de simples pauses de souffle qui participent au rythme de la parole ;
- Les pauses moyennes [//] : ce sont des pauses marquées et décidées par l'auteur pour marquer frontières syntaxiques et séparer les différents énoncés du texte ;
- Les pauses longues [///] : ce sont des pauses qui segmentent et délimitent à coté des applaudissements [A] les paragraphes du texte. C'est ce type de pauses qui jouent le rôle de ponctuants des grandes unités textuelles (Adam, 2006 : 56). Les pauses longues sont marquées, soit par la décision de l'auteur, soit par la pression de l'assistance.

3.2. Les niveaux d'annotation

Contrairement aux différents niveaux d'annotations exposées dans le cadre théorique, dans notre application nous avons porté quelques modifications. En effet, nous avons opté pour quatre niveaux principaux :

- Une balise ***TEXT*** qui prend en charge l'intégralité du texte ;
- Une balise ***S*** qui permet de délimiter les différents paragraphes du texte ;
- Une balise ***TRANSL*** qui permet de traduire en français les différents paragraphes ;
- Une balise ***E*** qui permet de représenter les différents énoncés figurant dans chaque paragraphe.

3.3. Présentation de l'application

Dans ce point, nous présentons les différentes balises compensant l'application de transcription synchronisée de l'échantillon de notre corpus.

3.3.1. Le fichier son

Comme nous l'avons signalé dans la partie théorique, le fichier son a été segmenté en différents niveaux textuels par le logiciel *Audacity*[3] . Chaque segment est caractérisé par un intervalle de durée représentant les temps initial et final (*cf*. Figure 1).

3 Le logiciel est téléchargeable sur le site : http://www.commentcamarche.net/download/telecharger-113-audacity

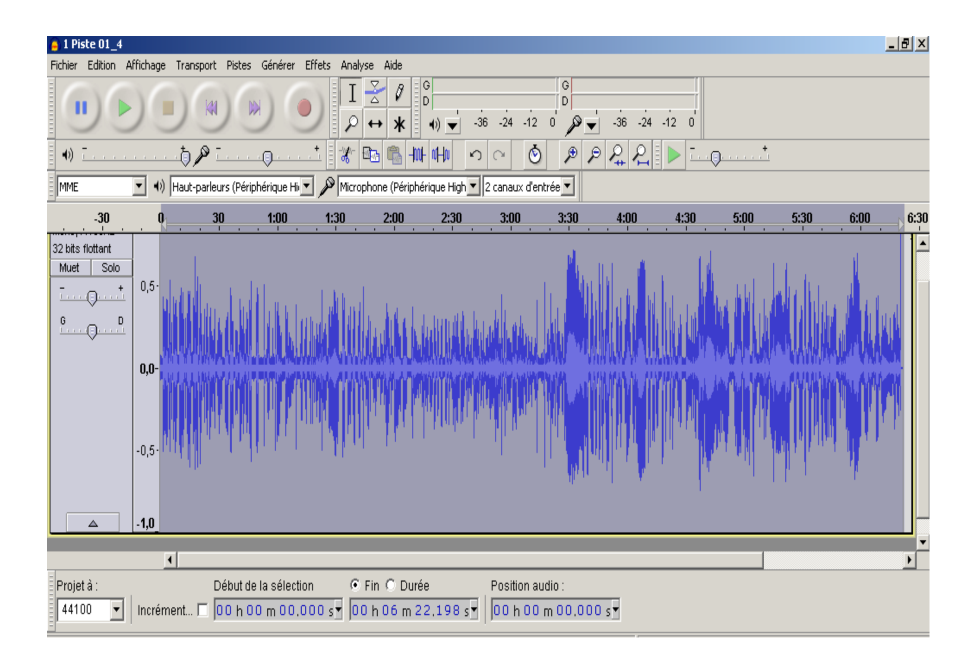


Figure 1 : Segmentation du fichier son à l'aide du logiciel Audacity.

3.3.2. La page d'accueil de l'application

Dans l'application, le fichier texte est représenté par les quatre niveaux d'annotation : Texte, paragraphe, traduction et énoncés avec la prise en considération des différentes pauses et les applaudissements qui figurent dans le corpus. Dans la première page d'accueil de l'application, la première ligne nous renseigne sur le titre du corpus et sa langue, suivie du ou/et des chercheurs qui ont transcrit le texte.

La troisième ligne représente « le fichier son » et les différents niveaux d'annotation du corpus. La dernière ligne nous renseigne sur les langues en contact des unités qui figure dans le corpus (*cf*. Figure 2).

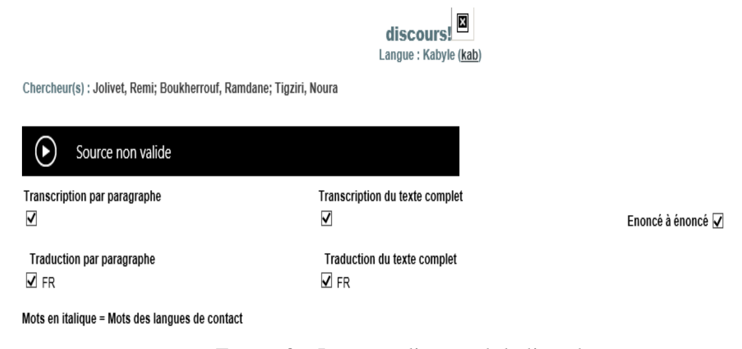


Figure 2 : La page d'accueil de l'application

3.3.3. La transcription synchronisée

La deuxième page de l'application comporte les différents niveaux de l'application. Le premier niveau représente le texte intégral, le deuxième concerne les différents paragraphes, le troisième est réservé à la traduction[4] française du corpus. Quant au quatrième et dernier niveau, il concerne les différents énoncés [E] de chaque paragraphe avec la prise en charge des différentes pauses. [PC] pour les pauses courtes, [PM] pour les pauses moyennes et [PL] pour les pauses longues sans oublier [AP] pour les applaudissements (*cf*. Figure 3).

Figure 3 : Les différents niveaux du corpus synchronisé.

4 Nous avons opté pour une traduction plus au moins fidèle pour nous permettre de représenter les référents culturels de la langue source.

3.3.4. Fonctionnement de l'application

Après avoir présenté ci-dessus les différentes parties composant l'application, dans ce point nous tenterons d'exposer son fonctionnement.

Les différents niveaux de l'application comportent un fichier son et un fichier texte. Pour nous permettre de lire et d'écouter simultanément l'un des niveaux annotés, il suffit juste de cocher la case correspondante. En effet, chaque niveau est caractérisé par un bouton qui nous permet de lire et d'écouter simultanément (*cf*. Figure 3).

Conclusion

En guise de conclusion, nous pouvons dire que notre travail est vraiment à l'étape d'expérimentation et d'essai. Notre projet vise à mettre en place une banque de données numérique d'un grand corpus kabyle synchronisé (fichier son- Texte) représentant l'ensemble des régions de la Kabylie, voir d'autres régions berbérophones.

Nous comptons mettre à la disposition des chercheurs berbérisants une banque de données des corpus oraux avec l'association des fichiers sons pour leurs faciliter de mener leurs recherches dans l'ensemble des niveaux de la langue (phonétique, phonologique, morphologique, sémantique, syntaxique, textuel, etc.).

Bibliographie

Adam J-M. (2005). *La linguistique textuelle, introduction à l'analyse textuelle des discours*, Paris, Armand Colin, « Cursus ».

Blanche-Benveniste C. & Jeanjean C. (1987). *Le français parlé, Transcription et édition*. Paris, Didier Erudition.

Blanche-Benveniste C. (1997). *Approches de la langue parlée en français*, Paris, Ophrys.

Bouzouita S. (2008). Cohésion et hiérarchisation des unités intonatives et textuelles dans l'oral spontané, Thèse de doctorat, Université de Paris III, Sorbonne nouvelle.

Candea M. (2000). Contribution à l'étude des pauses silencieuses et des phénomènes dits «d'hésitation» en français oral spontané, Thèse de doctorat, Université de Paris III, Sorbonne nouvelle.

Chaker S. (1983). *Un parler berbère d'Algérie(Kabylie) Syntaxe*, Université de Provence.

Chaker S. (1991). *Manuel de linguistique berbère*, Bouchène, Alger.

Duez D. (1991). *La pause dans la parole de l'homme politique*, CNRS Editions.

Mettouchi A. (2006). Anaphoricité et appel à l'attention partagée dans un conte oral en kabyle (berbère», *Loquentes Linguis, Studi linguistici e orientali in onore di Fabrizio A. Pennacchietti*, P.-G. Borbone, A. Mengozzi & M. Tosco (eds), Wiesbaden : Harrassowicz, PP. 499-507.

Jolivet R. (2013). Annotation de corpus linguistique et TEI, *Colloque International : Linguistique du corpus : recueil, exploitation et diffusion*, 23 et 24 juin, laboratoire d'Aménagement et d'Enseignement de la Langue amazighe, Université Mouloud Mammeri de Tizi-Ouzou, Algérie.

Etude du comportement formantique statique et dynamique de consonnes labio-vélarisées spécifiques du kabyle algérien

Leila Falek **Hocine Teffahi** **Amar Djeradi**
USTHB. Faculté d'Electronique et d'Informatique. Alger
lfalek@hotmail.fr hteffahi@gmail.com adjeradi05@yahoo.com

Résumé

Le but de cette étude est une caractérisation acoustique de deux consonnes labio-vélarisées /gw/ et /kw/ spécifiques de la langue berbère, (plus particulièrement du kabyle algérien). Pour cela, nous avons utilisé les équations de locus et les trajectoires des formants. Nous avons proposé un procédé de représentation de la trajectoire du deuxième formant par la pente et l'ordonnée à l'origine de la régression linéaire. Cette méthode a permis de fournir des informations pour compléter les équations de locus et a aidé à faire la distinction entre les consonnes « labio-vélarisées » /gw/ et /kw/ et leurs homologues vélaires /k/ et /g/.

1. Introduction

Les dialectes berbères comptent des consonnes très spécifiques de leur région géographique. Ces consonnes sont parfois très difficiles à prononcer : seuls les natifs de la région arrivent à les produire. Elles découlent souvent de phonèmes déjà présents dans la majorité des langues du monde.

Dans cette étude, nous nous sommes intéressés à La labio-vélarisation des consonnes vélaires et labiales qui est un phénomène très répandu. Il est attesté dans tous les grands dialectes berbères du nord. Seul le touareg l'ignore totalement. Mais, à l'intérieur même des grandes régions berbérophones du nord, il est très instable : il peut être tout à fait absent dans certains parlers (comme la région de Bougie en Kabylie) et son extension et sa fréquence peuvent varier sensiblement entre les parlers qui le connaissent. Certains parlers de Grande Kabylie ou du domaine chleuh, par exemple, "labio-vélarisent" plus que d'autres de la même région. Ainsi, en Grande Kabylie, l'adjectif "grand" sera prononcé : amqqwran chez les At Yiraten et amqqran chez les At Yanni, distants de quelques kilomètres.

L'un des problèmes qui se posent à leur propos est de décider s'il s'agit d'un seul ou de deux phonèmes. Tous les critères utilisés par les linguistes dans ce genre de cas amènent à conclure qu'il s'agit de phonèmes uniques (un phonème et non succession de deux phonèmes, /k + w / par exemple). Les critères phonétiques et phonologiques sont par ailleurs nettement confirmés par les données puisque dans le système des alternances verbales, ces segments complexes sont traités comme des phonèmes uniques : agwem, "puiser, aller chercher de l'eau", akwer, "voler" ont exactement la même morphologie que asem "jalouser", arem "goûter", ames, "salir" ou afeg, "voler"... verbes à deux consonnes radicales (type aCC).

Au niveau de la notation, tant scientifique qu'usuelle, il faut en conséquence représenter les labio-vélarisées par une seule lettre, accompagnée, éventuellement, d'un diacritique (Chaker, 1996) :

- La représentation scientifique du type ne paraît pas adéquate pour l'usage courant car sa réalisation n'est aisée que sur micro-ordinateur.
- Le Fichier de Documentation Berbère a utilisé un petit ° (de n°) au-dessus de la ligne d'écriture : k°, g° ... qui a l'avantage d'être facile à réaliser puisqu'il fait partie des conventions typographiques générales dans les écritures latines et qu'il est présent sur un clavier de machine à écrire (Chaker, 1996).

L'écrivain Mouloud Mammeri, puis le groupe d'études berbères de Vincennes (revue Tisuraf) et une majorité des praticiens du kabyle des vingt dernières années, ont employé la succession "consonne + "semi-voyelle /w/" (CW) sur la même ligne d'écriture.

Dans cette étude, nous nous sommes intéressés à deux consonnes vélaires /g/ et /k/ labio-vélarisées : /gw/ et le /kw/ (nous avons adoptés l'ajout du « w ») pour la représentation du « ° » pour une facilité d'écriture en mode graphique. Il s'agit d'essayer de caractériser acoustiquement ces 4 consonnes et voir si on peut les distinguer (si ces vélaires se distinguent de leur homologue labio-vélarisées).

Cette caractérisation est faite sur la base de l'analyse des pentes des droites obtenues à partir des équations de locus et sur l'analyse dynamique des formants des voyelles suivant la consonne considérée.

2. Les équations de locus et les trajectoires formantiques

La production des sons de la parole nécessite le déplacement continuel des organes de la parole ; déplacement qui implique d'importantes modifications de la forme et du volume du conduit vocal. Ainsi, ces variations articulatoires conduisent acoustiquement à des variations rapides visibles sur les formants lorsqu'on produit une suite syllabique du type CV ou CVC, par exemple. Le passage d'une consonne à une voyelle par exemple va modifier la forme des formants vocaliques en produisant des petits changements appelés « transitions formantiques», qui peuvent être considérées comme des variations dynamiques rapides dans la structure des formants F1, F2, F3, F4, en fonction du temps.

Pierre Delattre et ses collaborateurs (Delattre *et al.*, 1952 ; Delattre *et al.*, 1955) ont proposé des expériences pour étudier les variations des transitions formantiques. À partir de stimuli synthétiques, ces chercheurs ont trouvé que les auditeurs étaient capables d'identifier le lieu d'articulation des consonnes occlusives orales voisées et non voisées et les consonnes nasales, lorsqu'un pattern de deux formants était associé à une transition formantique spécifique prenant origine à une valeur précise. Ils ont montré que lorsque les transitions formantiques de F2 dans une syllabe du type CV étaient montantes et ayant pour origine 720 Hz, la « meilleure » consonne /b/ était identifiée par les auditeurs ; lorsque la transition partait de 1800 Hz, la « meilleure » consonne /d/ est obtenue ; et pour la « meilleure » consonne /g/, la transition était descendante de la consonne à la voyelle pour atteindre les 3000 Hz (Delattre *et al.*, 1955), (voir Figure 1). Ils ont utilisé le terme « locus » pour rendre compte de cette valeur fixe, point virtuel origine des transitions.

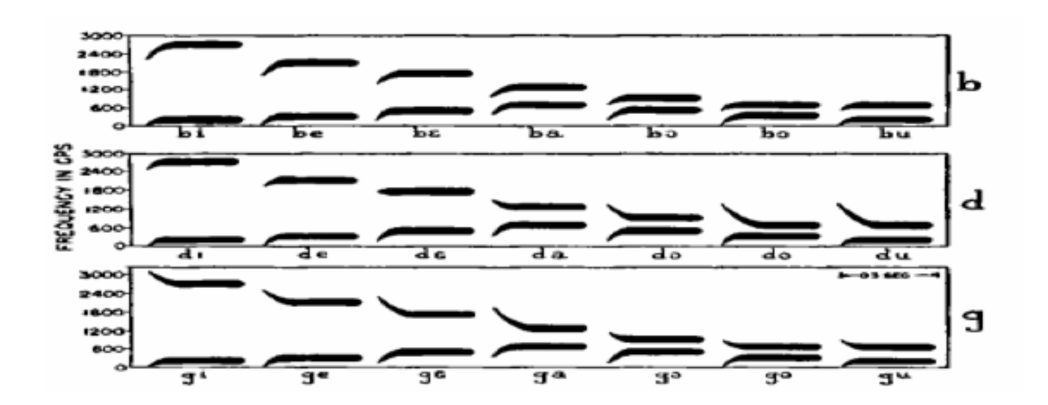


Figure 1 : Direction des transitions de F1 & de F2
synthétisées pour différentes voyelles montrant les lieux d'articulation
correspondant (Emprunté à Delattre et al., 1955)

On sait depuis, que ce sont les transitions formantiques de F2 et de F3, qui jouent un rôle déterminant pour indiquer le lieu d'articulation des consonnes adjacentes aux voyelles : ainsi, c'est la valeur du « locus virtuel » de F2 et la direction descendante ou montante de F3 qui donne une indication du lieu d'articulation. On sait également pour le lieu vélaire que ce sont deux valeurs de « locus virtuel » distinctes qui sont proposées en fonction de la voyelle adjacente : si la voyelle est antérieure, alors un « locus virtuel » autour des 3000 Hz (au maximum) est obtenu ; en revanche, si la voyelle est postérieure, alors un « locus virtuel » autour des 1300 Hz est obtenu (voir Figure 2) (Kewley-Port, 1982) et (Lindblom, 1990) entre autres.

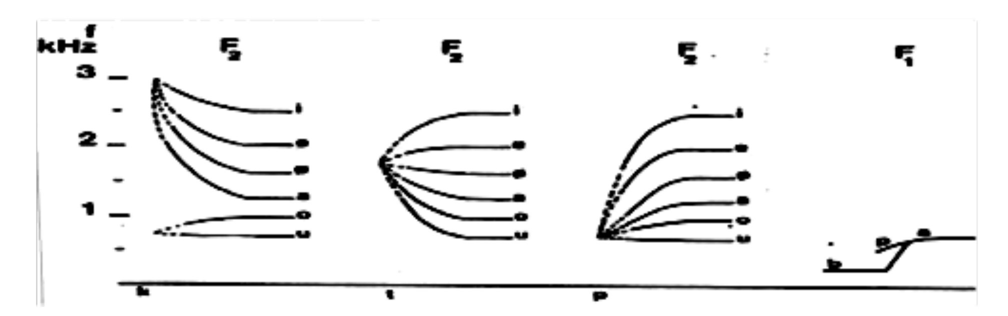


Figure 2 : Directions des transitions formantiques dans des syllabe CV avec la position du « locus virtuel » de F2 en fonction des consonnes /p t k/ et des voyelles /i e ø a o u/

3. Les équations de locus ou équations de lieu

Les équation de locus consistent en des régressions linéaires effectuées à partir de points représentatifs de la relation entre les Onset du deuxième formant de différentes voyelles (F2Onset) et leurs états stables correspondant (F2stable)(Chennoukh, 1995 ; Chennoukh *et al*., 1997).

Cette équation a la forme suivante :

$$F2Onset = m*F2stable+b \qquad (1)$$

(où m et b sont respectivement la pente et l'ordonnée à l'origine de l'équation de régression). Les valeurs de cette équation sont obtenues pour une seule consonne co-articulée avec différentes voyelles, où F2Onset représente l'ordonnée et F2stable l'abscisse.

Björn Lindblom (1963) est à l'origine du concept de l'équation de locus à partir d'une étude sur des séquences CV (où C = /b d g/ suivie des 8 voyelles du suédois) sur un seul sujet. Les coefficients des équations de locus obtenus varient en fonction du lieu d'articulation. Ainsi, il a trouvé que les pentes de la droite de régression variaient en fonction du lieu d'articulation de la consonne, de /g/ (0,95) > /b/ (0,69) > /d/ (0,28). L'auteur ne donne aucune interprétation de ces résultats mais explique que "Fitting strait lines serves only a descriptive purpose and is not motivated by deeper considerations as to the underlying complex articulatory processes" (Lindblom, 1963). L'auteur utilise la notion « locus » pour désigner une valeur mesurée concrète, à la place d'une valeur dérivée par le « locus virtuel » ; la valeur de F2Onset qui varie systématiquement avec la voyelle adjacente.

À partir de ces résultats, on peut facilement identifier le lieu d'articulation des consonnes, c'est-à-dire, le « locus virtuel », en fonction à la fois de la valeur de la pente de la droite de régression et de l'ordonnée à l'origine. La valeur du « locus virtuel » correspond au point d'intersection de la ligne de régression avec la diagonale à 45 degrés (ou en appliquant la formule suivante :

$$\text{locus virtuel} = b/(1-m)) \qquad (2)$$

Depuis l'étude de Lindblom (1963), plusieurs chercheurs ont exploité cette méthode afin d'identifier le lieu d'articulation des consonnes, nous pouvons citer entre autres : (Krull, 1987; Sussman *et al*., 1993 ; Sussman et Shore, 1996 ; Sussman et Modarresi, 2003; Fowler, 1994; Celdran et Villabla, 1995; Chennoukh *et al*., 1997; Yéou, 1995; Tabain et Butcher 1999; Modarresi *et al*., 2005), etc....

Toutefois, certains auteurs ont trouvé des différences pour les coefficients des équations de locus pour des consonnes ayant le même lieu d'articulation et différant par le mode d'articulation et/ou le voisement ((Fowler, 1994; Chennoukh *et al*., 1997; Yeou, 1997), etc…). Ils considèrent ainsi les coefficients obtenus comme de bons indicateurs du degré de coarticulation des voyelles avec les consonnes adjacentes : une valeur de pente de la droite de régression élevée (i.e., m = 1) indique une coarticulation maximale entre consonnes et voyelles (i.e., une résistance minimale du geste consonantique aux effets coarticulatoires des voyelles), tandis qu'une pente faible (i.e., m = 0) indique l'absence de coarticulation

entre consonnes et voyelles (i.e., une résistance maximale du geste consonantique aux effets coarticulatoires des voyelles).

La validité du concept de l'équation du locus a été confirmée dans plusieurs langues comme le thaï, l'urdu, et l'arabe égyptien (Sussman et Modarresi, 2003), le français, l'anglais américain et le suédois (Molis *et al*., 1994), l'anglais américain et le persan (Modarresi *et al*., 2005).

Tabain et Butcher (1999) ont expérimenté l'équation de locus dans la comparaison de deux langues aborigènes d'Australie très proches, le yanyuwa, le yindjibarndi, avec l'anglais australien. Les résultats de la littérature souvent similaires penchent pour une distinction nette entre deux groupes : un groupe de consonnes dentales et alvéolaires et un groupe de consonnes labiales et vélaires. Toutefois, les recherches indiquent que les pentes des droites de locus varient selon la langue (voir Tableau 1). Les natifs du même pays, voir de la même région géographique ont plus de chance de présenter des pentes similaires.

Les pentes des droites de locus sont aussi influencées par les caractères de voisement comme l'a souligné Sussman et Modarresi (2003) (Figure 3).

	Langue	/b/	/d/	/g/ (vélaire)
Lindblom (1963$_b$)	Suédois	0,69	0,28	0,95
Krull (1987)	Suédois	0,78	0,40	*
Sussman *et al.* (1991)	Anglais Américain	0,89	0,42	0,70
Sussman *et al.* (1993)	Thai	0,70	0,30	*
Sussman *et al.* (1993)	Arabe égyptien	0,77	0,25	0,92
Sussman *et al.* (1993)	Urdu	0,81	0,50	0,97
Fowler (1994)	Anglais Américain	0,80	0,48	0,71
Celdran & Villabla (1995)	Espagnol	0,83	0,58	1,00
Sussman & Shore (1996)	Anglais Américain	*	0,40	*
Chennoukh *et al.* (1997)	Ph1 : coarticulation maximale	0,85	0,51	0,98
Chennoukh *et al.* (1997)	Ph2 : coarticulation moyenne	0,75	0,42	0,86
Chennoukh *et al.* (1997)	Ph3 : coarticulation minimale	0,33	0,22	0,45
Sussman *et al.* (1997)	Anglais Américain	0,77	0,37	1,12
Yéou (1997)	Arabe Standard	0,92 (/f/)	0,48	0,90 (/χ/)
Tabain & Butcher (1999)	Yanyuwa	0,80 (/p/)	0,62 (/t/)	1,03 (/k/)
Tabain & Butcher (1999)	Yindjibarndi	0,83 (/p/)	0,48 (/t/)	1,11 (/k/)
Tabain (2000)	Anglais Australien	*	0,42	0,84
Sussman & Modarresi (2003)	Anglais Américain	0,80	0,41	0,89
Sussman & Modarresi (2003)	Perse	0,87	0,48	1,03
Al-Tamimi (2004)	Arabe Jordanien	0,63	0,36	0,87
Modarresi *et al.* (2005)	Anglais Américain	0,69 (@burst)	0,29 (@burst)	0,94 (@burst)
Modarresi *et al.* (2005)	Perse	0,80 (@burst)	0,30 (@burst)	1,00 (@burst)

Tableau 1 : Valeurs moyennes de pentes de la droite de régression des équations de locus, en fonction des consonnes /b/, /d/ et /g/, en contexte CV, à partir de différentes études. (= donnée non disponible ou contexte non traité). Si le contexte consonantique est différent, nous l'avons marqué entre parenthèse.*

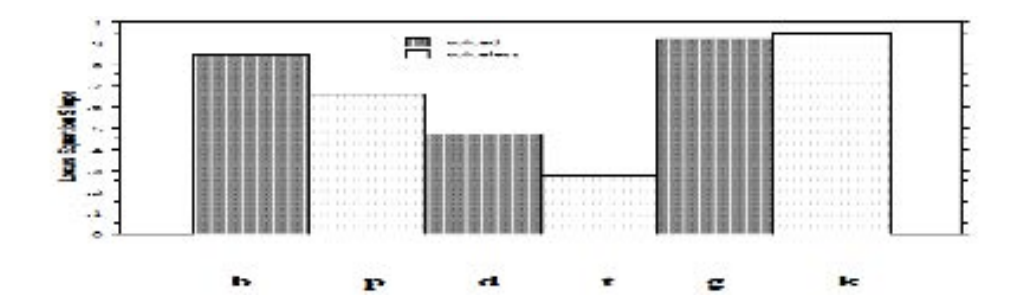


Figure 3 : Pentes des droites de locus en fonction du voisement des consonnes

3.1. Méthode d'extraction des Onset du deuxième formant de différentes voyelles (F2Onset) et leurs états stables F2stable

L'extraction du F2onset diffère selon que nous avons une plosive ou une consonne simple. La plosive est caractérisée par un burst (barre d'explosion). Ce qui engendre 2 valeurs du 2ème formant (voir Figure 4): F2onset@burst est une mesure effectuée directement après la barre d'explosion (Modarresi *et al*., 2005).

F2onset : Valeur de F2 prise 5 ms après le Burst (pour les plosives).

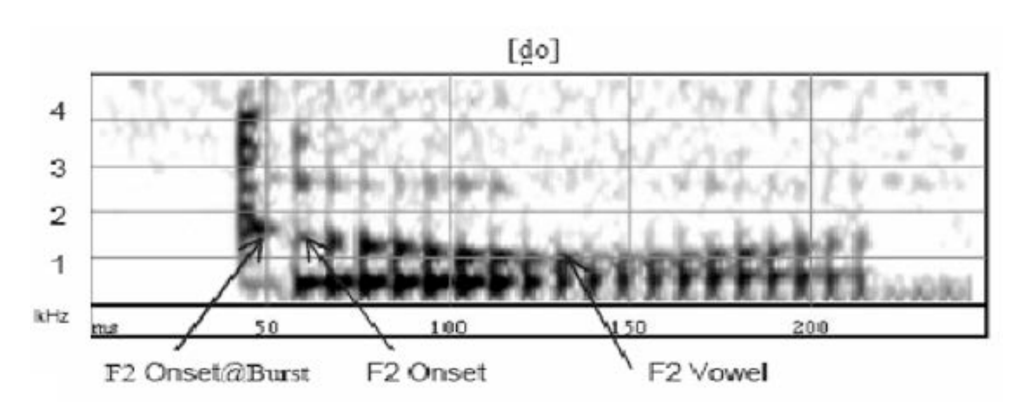


Figure 4 : Schéma présentant les méthodes d'extraction de l'onset : a) @Burst et b) à la première pulsation après le relâchement de la voyelle

Dans cette étude nous avons adopté la valeur de F2onset : valeur de F2 prise 5 ms après le Burst.

4. Analyse dynamique par les trajectoires formantiques et équations de locus

Par analyse dynamique nous entendons une analyse du comportement des formants des voyelles en prenant en compte plusieurs valeurs formantiques au cours de la réalisation des voyelles à droite de la consonne considérée, en comparaison avec l'utilisation de deux valeurs correspondant au début de la transition consonne/voyelle (F2onset) et la valeur F2stable : position centrale, utilisées dans l'analyse par les équations de locus.

La différence entre les deux méthodes est que les équations de locus, qui sont obtenues à partir de deux valeurs formantiques obtenues à F2Onset et à F2Milieu, donnent une bonne description du lieu d'articulation et/ou du degré de coarticulation ; tandis que les pentes formantiques qui sont obtenues à partir de plusieurs valeurs formantiques (au minimum cinq valeurs pour une durée de transition de 20 ms), rendent compte du degré de coarticulation et du lieu d'articulation des consonnes adjacentes pour chaque voyelle et chaque formant.

La méthode de calcul des pentes formantiques est empruntée à EL-Tamimi (2007).

La pente formantique calculée ici est plus proche de la vitesse de transition formantique. Comme nous l'avons expliqué, l'onset vocalique est mesuré à 5 ms après le relâchement de la voyelle, position située toujours au milieu de la première période de la voyelle.

Dès que l'onset est déterminé, toutes les valeurs obtenues à 5 ms d'intervalle sont utilisées pour effectuer le calcul des pentes formantiques.

Ainsi, et pour tous les formants, les pentes formantiques sont obtenues par une analyse de régression linéaire en partant de la première valeur formantique, obtenue à 5 ms du début, qui est placée au temps 0, jusqu'à la valeur au milieu temporel de la voyelle.

Nous appellerons l'intervalle de temps entre le point 0 et le milieu temporel la « durée de transition », qu'il ne faut cependant pas confondre avec la terminologie traditionnelle. La durée de la transition est traditionnellement quantifiée à partir du point 0 (début de la voyelle) jusqu'à la fin des changements abruptes dus au passage de la consonne à la voyelle.

Kent et Moll (1969) expliquent que la durée de la transition vers la voyelle en contexte CV est un invariant dynamique en production des occlusives voisées et non-voisées et correspond à une durée proche des 50 ms. Au-delà de ces 50 ms, l'état stable de la voyelle débute.

Dans notre cas, nous prenons en compte la durée de la transition traditionnelle, mais également une partie de la durée de l'état stable. Pour toutes les voyelles ayant une durée vocalique inférieure à 100 ms, la durée de la transition traditionnelle et celle que nous utilisons seront les mêmes, à partir des suppositions de Kent & Moll (1969). Pour les voyelles ayant une durée vocalique supérieure à 100 ms, la transition traditionnelle est prise en compte avec en plus une partie de l'état stable.

5. Partie expérimentale

5.1. Les unités linguistiques utilisées

Nous avons choisi de caractériser la labio-vélarisation de deux consonnes vélaires /g/ (voisée) et /k/ (non voisée).

L'un des outils de caractérisation est l'équation de locus. Les valeurs de cette équation sont obtenues pour une seule consonne co-articulée avec différentes voyelles. Ce qui exige donc un matériau linguistique composé de ViCVj (Vi et Vj : voyelles berbères /a/, /i/, /u/ et /C/ : /g/, /k/, /gw/, /kw/). Le nombre de locuteurs était de 4 (AM, AN TH et FL), tous natifs de la grande Kabylie et vivant à Alger. Chaque locuteur a subi une phase d'apprentissage du corpus avant l'enregistrement (pour bien prononcer les consonnes ciblées). Il lui est demandé de répéter 10 fois chaque ViCVj (pour avoir une moyenne de chaque caractéristique). L'enregistrement a eu lieu dans la chambre sourde du laboratoire d'acoustique de la Faculté d'Electronique et d'Informatique de l'USTHB.

Le corpus est ainsi composé de quatre consonnes, insérées dans des ViCVj, répétées 10 fois chacune, avec changement de contexte vocalique /a/, /i/, /u/, droit et gauche de la consonne, avec un nombre de locuteurs de deux. Ce qui fait un corpus de : 4*10 *9*4=1440 unités d'analyse.

Les Tableaux 2 illustrent les différents ViCVj pour chaque consonne /C/ :

Vowel	Consonant /g/		
/a/	aga	agi	agu
/i/	iga	igi	igu
/u/	uga	ugi	ugu

Vowel	Consonant /k/		
/a/	aka	aki	aku
/i/	ika	iki	iku
/u/	uka	uki	uku

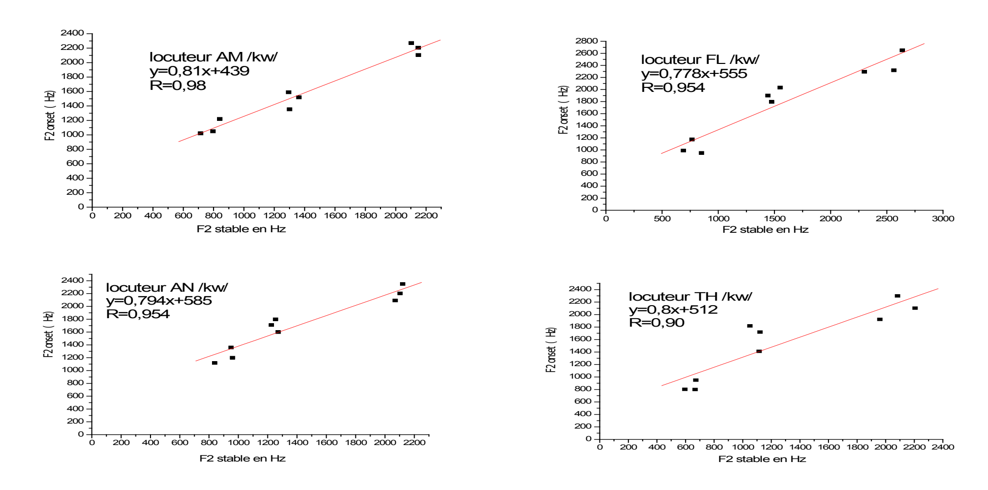
Vowel	Consonant /gw/		
/a/	agwa	agwi	agwu
/i/	igwa	igwi	igwu
/u/	uga	ugui	ugu

Vowel	Consonant /kw/		
/a/	akwa	akwi	akwu
/i/	ikwa	ikwi	ikwu
/u/	ukwa	ukwi	ukwu

Tableaux 2 : Les unités linguistiques ViCVj pour chaque consonne /C/ utilisée

5.2. Résultats obtenus pour les équations de locus

Nous avons tracé les équations de locus F2onset (F2stable) pour les quatre consonnes /g/, /gw/, /k/ et /kw/ et pour les quatre locuteurs en considérant la moyenne de toutes les fréquences F2onset et F2stable des 10 répétitions des unités VCV enregistrées pour chaque locuteur. Un exemple de résultats obtenus pour les droites de locus pour la consonne /kw/, pour chaque locuteur, est illustré par les Figures 5.



Figures 5 : Droites de locus F2onset (F2stable) pour la consonne /kw/ pour les quatre locuteurs

Les pentes m et ordonnées à l'origine b des droites de locus obtenues, suivies de leurs coefficients de régression R sont regroupées dans le Tableau 3.

La relation (3) rappelle l'expression de l'équation de locus

$$F2Onset = m*F2stable+b \qquad (3)$$

Pour les droites de locus obtenues par consonne et par locuteur.

Consonne	/g/			/gw/			/k/			/kw/		
Locuteur	**m**	**b**	**R**	**m**	**b**	**R**	**m**	**b**	**R**	**m**	**b**	**R**
AM	0.86	469	0.96	0.71	465	0.96	1.05	332,37	0.97	0.81	439	0.98
FL	0.89	411	0.99	0.628	778	0.94	0,93	594	0.97	0.77	555	0.95
AN	0.809	411.7	0.99	0.68	728	0.94	0,94	500	0.97	0.79	585	0.95
TH	0.799	485	0.9	0.74	421	0.93	0,998	582	0.99	0.8	512	0.90

Tableau 3 : Résultats obtenus pour les équations de locus (pente m, ordonnée à l'origine b et coefficients de régression R) par consonne et par locuteur

Les résultats du Tableau 3 montrent des valeurs de pentes de droites de locus distinctes pour les consonnes /g/ et /k/ qui rejoignent celles obtenues dans la littérature pour la langue arabe.

Les valeurs trouvées pour les pentes des droites de locus de /g/ et /k/ rejoignent celles obtenues dans la littérature par Sussman et Modarresi (2003). Ces derniers affirment que la différence entre la pente de /g/ et celle de /k/ est liée au caractère voisé (/g/) et non voisé (/k/) des deux consonnes comme le montre la Figure 3. Mohamed Embarki *et al*. (2007 ; 2010) ont montré que la valeur de la pente est liée aussi à la langue.

Nous avons noté aussi une diminution des pentes des droites de locus pour les consonnes labio-vélarisées /gw/ et /kw/, telles que: m/k/> m/kw/ et m/g/> m/gw/. Comme la valeur de la pente des droites de locus augmente avec la coarticulation de la consonne avec la voyelle, nous dirons donc que la labio-vélarisation de ces deux consonnes provoque une diminution de la coarticulation.

5.2. Résultats obtenus pour les pentes des trajectoires formantiques du deuxième formant

Nous avons déterminé les pentes et les ordonnées à l'origine des régressions linéaires des transitions formantiques du formant F2 (depuis le début du voisement jusqu'à la stabilité du F2 comme dans le cas de la détermination des équations de locus) pour chaque voyelle /a/, /i/ et /u/ du contexte droit de chaque consonne /k/ /kw/ et /g/ et /gw/ (pour les 10 répétitions effectuées). Nous avons ensuite tracé les ellipses de dispersions des couples (pentes, ordonnées à l'origine) pour les trois locuteurs AM, AN et TH, pour chaque consonne avec son homologue labio-vélarisé, dans un même graphe pour 3 locuteurs.

La Figure 6 montre un exemple de régressions linéaires obtenues pour les trajectoires formantiques pour une voyelle /a/ du contexte droit de l'unité /aga/, suivie de leurs équations pour chaque formant

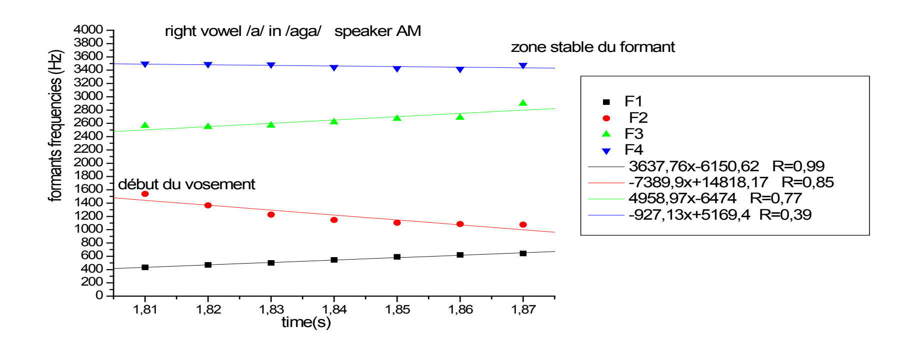
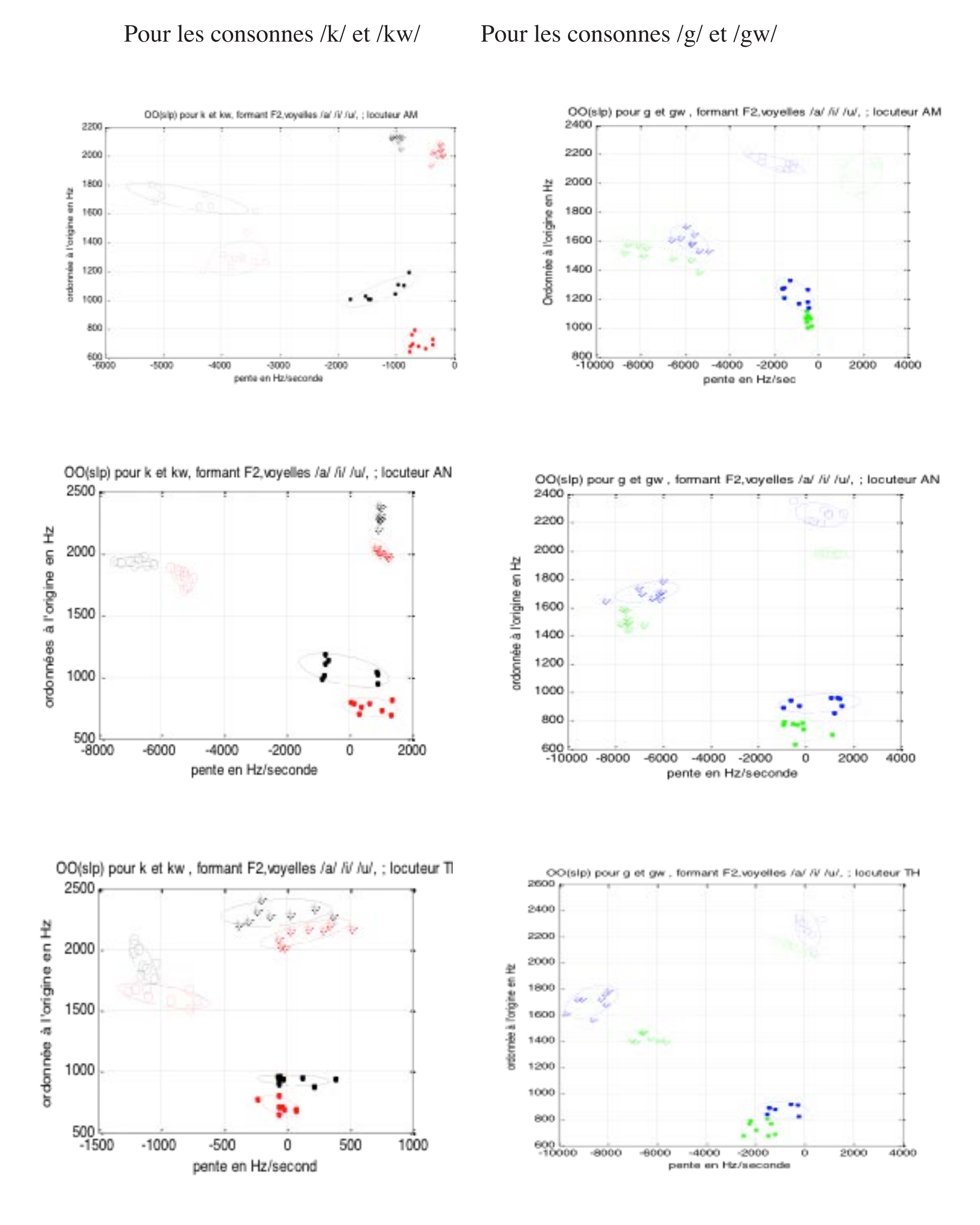


Figure 6 : Exemple de régressions linéaires avec leurs équations

Les résultats obtenus sont illustrés sur la Figure 6.



Figures 7 : Ordonnées à l'origine des transitions formantique du 2ème formants pour les voyelles /a/, /i/, /u/ du contexte droit des consonnes /k/, /kw/, /g/ et /gw/ : /k/en rouge, /kw/ en noir, /g/ en vert, /gw/ en bleu

Les Figures 7 montrent une augmentation des ordonnées à l'origine pour les consonnes labio-vélarisées, donc les F2onset sont plus élevés. Cela peut impliquer que la labio-vélarisation provoque une augmentation du F2. Par conséquent, nous pouvons en déduire que les mouvements de la langue (l'avancement de la langue) sont plus importants et donc plus d'antériorité en présence des consonnes labio-vélarisées.

Cependant, les pentes de ces transitions sont indifférentes pour la labio-vélarisation. Notons quand même que les pentes sont négatives et sont différentes pour les couples /k/ et/kw/ et /g/ et /gw/ en présence de la voyelle /a/ en contexte vocalique droit. La pente accuse alors une grande augmentation (en valeur absolue) pour le couple /g/ et /gw/, donc le F2onset augmente dans ce cas. Ce qui pourrait être un paramètre de distinction entre les deux couples.

Conclusion

Cette étude montre que la labio-vélarisation permet d'obtenir de nouvelles consonnes puisque les trajectoires des F2 sont différentes. De plus, les équations de locus ont montré des pentes différentes et en diminution lors de la labio-vélarisation.

L'utilisation des pentes et ordonnées à l'origine des régressions linéaires des trajectoires formantiques (limitées par le F2onset et le F2stable comme dans le cas du tracé des droites de locus) a permis d'apporter une information supplémentaire par rapport à l'antériorité de la langue qui accuse une augmentation lors de la labio-vélarisation.

Les pentes de ces régressions permettent aussi de faire la distinction entre les deux couples.

Références

Celdran E. M., Villalba X. (1995). Locus equations as a metric for place of articulation in automatic speech recognition. *The XIIIth International Congress of Phonetic Sciences*, Vol.1, pp.30-33.

Chaker S. (1996). Propositions pour la notation usuelle à base latine du berbère. *Synthèse des travaux de l'atelier « Problèmes en suspens de la notation usuelle à base latine du berbère »*, 24-25 juin, 1996.

Chennoukh S. (1995). Modélisation du conduit vocal en régions distinctives. Synthèse d'ensembles voyelle-voyelle et voyelle-consonne-voyelle. Thèse de doctorat, Ecole nationale supérieure des télécommunications de Paris.

Chennoukh S., Carré R., Lindblom B. (1997). Locus equation in the light of articulatory modeling. *Acoustical Society of America*, Vol. 102, n° 4.

Delattre P., Liberman A. M., Cooper F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 27(4): 769-773.

Delattre P., Liberman A. M., Cooper F. S., Gerstman L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one and two-formant vowels synthesized from spectrographic patterns. *Word*, Vol. 8, pp. 195-210.

El-Tamimi J. (2007). Indices dynamiques et perception des voyelles : Étude translinguistique en arabe dialectal et en français. Thèse de Doctorat, École Doctorale Humanité et Sciences Humaines, Université Lumière Lyon 2.

Embarki M., Ahmad A. (2010). L'équation de locus comme mesure de distinction sociale de genre en arabe koweïtien. *XXVIIIème Journées d'Etudes sur la Parole*, Mons, 25-28 mai, 2010.

Embarki M., Guilleminot Ch., Yeou M. (2006). Equation de locus comme indice de distinction consonantique pharyngalisé vs non pharyngalisé en arabe. *XXVIème Journées d'Etudes sur la Parole*, pp. 155-158.

Fowler C. A. (1994). Invariants, Specifiers, Cues: An Investigation of Locus Equations as Information for Place of Articulation. *Perception & Psychophysics*, 55(6): 597-610.

Kent R. D., Moll K. L. (1969). Vocal-tract characteristics of the stop cognates. *Journal of the Acoustical Society of America*, Vol. 46, pp. 1549-1555.

Kewley-Port D. (1982). Measurments of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, 72(2): 379-389.

Krull D. (1987). Second Formant Locus patterns as a measure of Consonant-Vowel Coarticulation. *Phonetic Experimental Research at the Institute of Linguistics*, University of Stockholm, Vol. 5, pp. 43-61.

Landercy A., Renard R. (1977). Élements de phonétique. Didier, Bruxelles.

Lindblom B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, Vol. 35, pp. 1773-1781.

Lindblom B. (1990). Explaining phonetic variation: A sketch of the H and H theory. *Speech production and speech modelling*, Hardcastle W. J., Marchal A. (Eds), Kluwer Academic Publishers, Dordrecht, pp. 403-439.

Modarresi G., Sussman H. M. (2003). The effect of syllable structure on bidirectional coarticulation. *Journal of the Acoustical Society of America*, 27(4): 3197-3200.

Modarresi G., Sussman H. M., Lindblom B., Burlingame E. (2005). Letter to the Editor: Locus Equation Encoding of Stop Place: Revisiting the Voicing/VOT Issue. *Journal of Phonetics*, Vol. 33, pp. 101-113.

Molis M.R., Lindblom B., Castelman W., Carré R. (1994). Cross-language analysis of VCV coarticulation. *Journal of the Acoustical Society of America*, Vol.95.

Sussman H.M., Hoemeke K., Ahmed F. (1993). A cross-linguistic investigation of locus equations as a relationally invariant descriptor of place of articulation. *Journal of the Acoustical Society of America*, Vol. 94, pp. 1256-1268.

Sussman H. M., Modarresi G. (2003). The stability of locus equation encoding of stop place. *The 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 3-9 August, 2003.

Sussman H. M., Shore J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception & Psychophysics*, 58 (6): 936-946.

Tabain M., Butcher A. (1999). Slope values as acoustic measures of coarticulation: a cross-language comparison of stop consonants. *Journal of Phonetics*, Vol. 27, pp.333-357.

Yeou M. (1997). Locus equations and the degree of coarticulation of Arabic consonants. *Phonetica*, Vol. 54, pp. 187-202.

Système d'identification de locuteurs en mode dépendant et indépendant de texte, appliqué à la langue arabe et berbère

Fatma zohra Chelali Khadidja Saddedine
Amar Djeradi Hocine Teffahi
Laboratoire de Communication parlée et traitement du signal, Faculté d'électronique et d'informatique, Université des sciences et de technologie Houari Boumedienne, Alger,
Chelali_zohra@yahoo.fr sadeddine_khadidja@yahoo.com
adjeradi05@yahoo.com hteffahi@gmail.com

Résumé

Notre travail consiste à établir un système d'identification de locuteurs en mode dépendant et indépendant de texte pour deux bases de données acoustiques l'une pour la langue arabe alors que la deuxième est pour la langue berbère. Pour cela, notre choix s'est porté sur les coefficients issus de l'analyse prédictive linéaire (LPC), leurs dérivées premières et secondes et la fréquence fondamentale. Le classifieur utilisé se base sur le maximum de vraisemblance. Dés résultats satisfaisants ont été obtenus en incluant les dérivées premières et secondes aux coefficients LPC.

1. Introduction

La reconnaissance automatique de la parole et du locuteur connait ces dernières années un grand intérêt de la part de beaucoup de chercheurs dans la conception des systèmes capables de reconnaître des sons ou d'identifier des locuteurs. Nous distinguons la reconnaissance du locuteur indépendante du contenu de la phrase prononcée (mode indépendant au texte) et la reconnaissance du locuteur qui prononce un mot ou une phrase clef (mode dépendant du texte) (Mami, 2003).

Dans cet article, nous nous intéressons essentiellement à la reconnaissance du locuteur en mode dépendant et indépendant du texte appliqué à la langue arabe et berbère.

Pour cela, un corpus acoustique a été enregistré au niveau du laboratoire (laboratoire communication parlée et traitement du signal) pour des phrases en parole continue de la langue berbère. De plus, une base de données de mots isolés comportant des fricatives et des emphatiques de l'arabe standard pour 6 locuteurs a été utilisée.

Plusieurs études ont été présentées pour beaucoup de langues, mais les particularités de la langue arabe et berbère en raison de ces traits acoustiques spécifiques comme l'emphase qui caractérise le consonantisme du berbère posent de nombreux problèmes aux constructeurs des systèmes de reconnaissance automatique de la parole (RAP).

Saeed et Kheir Nammoud (2007) dans leur système d'identification de locuteur en mode dépendant du texte pour les digits de zéros à dix en langue arabe achèvent un taux de reconnaissance de 98.8% en utilisant le modèle d'estimation de Burg. (Tanprasert *et al.*, 2000) dans leur travail présente un système d'identification de locuteur dépendant du texte appliqué à la langue thaïlandaise en utilisant le réseau de neurone dont les paramètres caractéristiques sont les coefficients LPC aboutissent à un taux de reconnaissance de 95%. Toutios et Margaritis (2002) ont extrait, pour le développement d'un système OGI TOOLKIT d'identification de locuteur en mode dépendant du texte, les coefficients MFCC et les paramètres PLP. En utilisant un réseau de neurone, ils ont obtenu un taux de reconnaissance de 93.10% pour la première base en langue anglaise.

Notre travail consiste donc à établir un système d'identification de locuteurs en mode dépendant et indépendant de texte pour les deux bases de données citées auparavant (arabe et berbères). Pour cela, notre choix s'est porté sur les coefficients issus de l'analyse prédictive linéaire (LPC) et la fréquence fondamentale. Le classifieur utilisé se base sur le maximum de vraisemblance.

La première Section donnera une brève description des caractéristiques phonétiques et phonologiques de la langue arabe et berbère. La seconde Section est consacrée à la présentation du corpus d'enregistrement, à la détermination des vecteurs descripteurs (coefficients LPC, la première et la seconde dérivée de ces coefficients) ainsi que les résultats obtenus lors de l'identification. Une reconnaissance de locuteurs par fréquence fondamentale est présentée dans la troisième Section. Nous terminerons par une conclusion qui résume les résultats obtenus et les perspectives d'avenir.

2. Phonologie arabe et berbère

2.1. Langue arabe

La classification phonétique se distingue en classes et sous classes, chaque classe étant liée à un mode articulatoire de l'appareil phonatoire. On distingue les consonnes et les voyelles.

Nous avons une voyelle lorsque le mode de production est caractérisé par le libre passage de l'air dans les cavités situées au-dessus de la glotte. La plupart des voyelles utilisées dans les langues sont sonores, c'est-à-dire qu'elles sont prononcées avec une vibration des cordes vocales au contraire des voyelles sourdes qui sont prononcées sans vibration des cordes vocales.

Les voyelles sont opposées aux consonnes, car ces dernières se caractérisent par une obstruction au passage de l'air. D'un point de vue perceptif, les voyelles se manifestent par des sons « clairs » tandis que les consonnes se caractérisent par des bruits tels qu'un, sifflement, un roulement, un claquement, etc.

Il existe des consonnes et des voyelles communes à toutes les langues par contre certaines sont spécifiques à une seule langue. Nous allons présenter quelques propriétés de la langue arabe et berbère.

L'arabe classique standard comporte 28 consonnes, 6 voyelles (3 longues et 3 courtes) et quelques autres réalisations vocaliques (gémination, silence).

Les 28 consonnes arabes sont divisées en deux groupes ; 14 consonnes solaires qui assimilent le « ل » de l'article et 14 consonnes lunaires qui n'assimilent pas le « ل » de l'article (Saidane *et al*., 2004).

L'arabe est une langue à consonantisme riche : 26 phonèmes consonantiques sans compter le / w/ الواو et le / y/ الياء. Ces consonnes peuvent être classées en fonction de leur mode d'articulation et d'émission (Hasanat, 2007).

Les consonnes arabes sont classifiées comme orales non emphatiques, emphatiques, liquides et nasales. La langue arabe est caractérisée par certaines consonnes qui n'existent pas dans toutes les langues. En plus de ces consonnes, l'arabe possède quatre autres consonnes emphatiques :

[đ], **[ṭ]**, **[ş]** et **[ð]**.

L'arabe peut renforcer une consonne quelconque. Ce renforcement est indiqué à l'aide du signe" ّ "« shadda ». Ce terme qui correspond à la « gémination » n'apparaît pas dans le tableau du mode d'articulation des consonnes. Ce signe prolonge la durée d'une articulation consonantique (environ de moitié et en augmentant l'intensité) (Hasanat, 2007).

Une étude sur l'analyse formantique ainsi que la durée des voyelles et des emphatiques berbères a déjà fait l'objet d'un article présenté à la conférence TICAM 2012 (Chelali *et al*., 2012).

2.2. Langue berbère

La langue berbère, ou tamazight, est présente à l'heure actuelle dans une dizaine de pays de l'ensemble Maghreb-Sahara-Sahel : Maroc, Algérie, Tunisie, Libye, Égypte, Niger, Mali, Burkina-Faso et Mauritanie. Mais l'Algérie et le Maroc sont, de loin, les deux pays qui comptent les populations berbérophones les plus importantes. La renaissance volontariste de l'alphabet traditionnel, le tifinagh, a permis de suppléer à la mémoire collective, de traduire les œuvres majeures du patrimoine mondial et développer une littérature berbère qui répond à une forte demande. L'évaluation du nombre de berbérophones est une question difficile et controversée car il n'existe pas dans les pays concernés de recensements linguistiques systématiques et fiables. Si l'on soumet à la critique l'ensemble des chiffres avancés par les diverses sources, depuis les débuts de la présence coloniale française jusqu'à nos jours, on peut raisonnablement estimer les berbérophones à un pourcentage d'environ 25 % de la population en Algérie, et 40 % de la population au Maroc (Chaker, 2004).

Les consonnes kabyles peuvent être classifiées comme suit : spirantes et occlusives, emphatiques, labio vélarisées, etc. Le kabyle, et le berbère en général, comporte trois voyelles, plus une voyelle de lecture :

a est moins ouvert qu'en français, entre le « a » et le « e » [æ] ;

i se prononce entre le « i » [i] et le « é » français [e] ;

u se prononce « ou » [u].

La voyelle neutre « e » (voyelle-zéro ou schwa) apparaît pour éviter la constitution de groupes de plus de deux ou trois consonnes.

2.3. Langue arabe/ Langue berbère

Après l'étude des caractéristiques phonologiques et phonétiques des deux langues arabe et berbère nous relevons les différences suivantes:

- La langue arabe s'écrit de droite à gauche contrairement à langue kabyle.
- Les lettres arabes ont plusieurs réalisations graphiques par contre les lettres kabyle n'ont qu'une seule.
- Selon les chercheurs, les consonnes épaisses dans la langue kabyle ne présentent pas la gémination mais une tension, par contre, les consonnes arabes peuvent être géminées (Nait-Zerrad, 1998).
- La voyelle neutre « e » existe dans le vocalisme kabyle est absente dans l'arabe.

3. Résultats expérimentaux

Nous présentons dans ce qui suit les résultats expérimentaux relatifs à l'identification de locuteurs en mode dépendant et indépendant de texte. Nous commencerons à présenter la base de données que nous avons enregistré au niveau du laboratoire LCPTS, puis les tests réalisés dans les deux modes pour établir un système d'identification de locuteurs en se basant sur les coefficients issus de l'analyse LPC ainsi que leurs dérivées premières et secondes et la valeur de la fréquence fondamentale. Une comparaison entre les résultats obtenus pour la langue arabe et langue berbère est ainsi faite.

L'identification et la vérification automatique du locuteur sont les tâches pionnières du domaine de la RAL regroupant plusieurs chercheurs dans ce domaine: plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'indexation par locuteur de flux audio, ou le suivi de locuteurs ou de nouvelles variantes telles que la détection du locuteur dans une conversation.

3.1. Corpus acoustique

Nous avons choisi six locuteurs pour la langue berbère (kabyle) qui sont : Nacera, Zohra, Ghania, Gassi, Kaci et Smail, tous natifs de la région de la Kabylie. Ces locuteurs répètent avec un débit moyen et une énergie moyenne six phrases en parole continue.

Une autre base de données prononcée par six locuteurs de la langue arabe: Chafik, Abidat, Amine, Chafika, Amira et Daouia établie auparavant a été utilisée à titre comparatif.

La carte son que nous avons utilisé est dénommée MobilePre USB. MobilePre USB est un préamplificateur mobile intégrant une interface audio parfaite pour les enregistrements sur ordinateur (portable ou mobile). En fait, il s'agit de l'un des seuls préamplis disponibles complètement alimenté via le bus USB.

Nous avons enregistré un corpus formé de six phrases berbères dont leurs durées varient entre 2s et 3s. Pendant l'enregistrement, chaque répétition a été réécoutée pour s'assurer que les phrases entières ont été correctement enregistrées et en évitant toute perturbation extérieure. Les phrases choisies sont comme suit :

« assagui yelha lhal », « azeka anrouh arthemourth », « anzoum remthane g nevdhou », «yetchak izeme yellozen », « yeghine wagroud aghour yemasse » et « thelha thazalith gakhem rebi ».

Les trois premières phrases sont de durée de deux secondes tandis que les trois dernières de durée égale à 3 secondes. La segmentation a été effectuée manuellement en utilisant le logiciel d'enregistrement Wavesurfer (sous formats wav) afin de prendre uniquement la partie essentielle où la parole a été produite, une normalisation des fichiers a été effectuée.

Après la segmentation, nous avons stocké les sons enregistrés comme suit :

Pour le mode dépendant du texte, nous avons pris la même phrase pour les dix répétitions des six locuteurs. Donc, on aura 60 fichiers son pour chaque phrase. Pour le mode indépendant du texte, nous avons pris les répétitions des six phrases pour un seul locuteur. Donc, un locuteur contient 40 fichiers son pour les 4 phrases de 2s et 20 fichiers son pour les phrases de 3s.

3.2. Calcul des paramètres de prédiction linéaire LPC

Il existe une diversité de méthodes pour extraire les caractéristiques d'un signal vocal, comme les coefficients spectraux et d'autres. Parmi les descripteurs les plus utilisés et simples à implémenter sont ceux issus du codage prédictif linéaire (LPC). Cette méthode permet de convertir l'information en vecteur contenant les caractéristiques les plus importantes.

Notre objectif est de calculer les vecteurs des coefficients LPC, caractérisant les signaux de parole, pour l'ensemble des répétitions et l'ensemble des locuteurs. Nous présentons dans ce qui suit les étapes essentielles au calcul des coefficients LPC. Nous avons normalisé les signaux segmentés de notre base de données pour avoir la même taille des fichiers son pour chaque phrases prise indépendamment.

Chaque signal vocal est décomposé en un nombre de trames. Pour chaque trame (ou fenêtre) nous déterminerons p coefficients LPC, ce qui va fournir un vecteur de $p*m$ coefficients tel que m est le nombre de trames pour chaque fichier acoustique. Les coefficients LPC sont calculés pour les nouveaux signaux normalisés et stockés dans une matrice de j lignes et i colones ($i=5$) pour chaque locuteur pour le mode dépendant du texte, et j lignes et i colones ($i=5$) pour chaque phrase pour le mode independant au texte. Nous rappelons que le $i=5$ représente les cinq répétitions pour chacune des matrices apprentissage et test.

Le module de décision, qui selon les critères établis, permet de designer le locuteur reconnu. Nous utiliserons dans notre travail la méthode de corrélation pour la reconnaissance.

Pour estimer la qualité de la reconnaissance de locuteur, il faut d'abord calculer le taux de reconnaissance (TR) dont le seuil du coefficient de corrélation (R) que nous avons choisi vaut 0.75.

$$TR\ (\%) = \frac{\text{Nombre de fichiers test tel que R > 0,75}}{\text{Nombre total}} * 100 \qquad (1)$$

Les histogrammes suivants indiquent les résultats de reconnaissance obtenus pour les phrases berbères et les mots arabes pour le mode dépendant du texte où l'ordre de prédiction varie de 10 à 18.

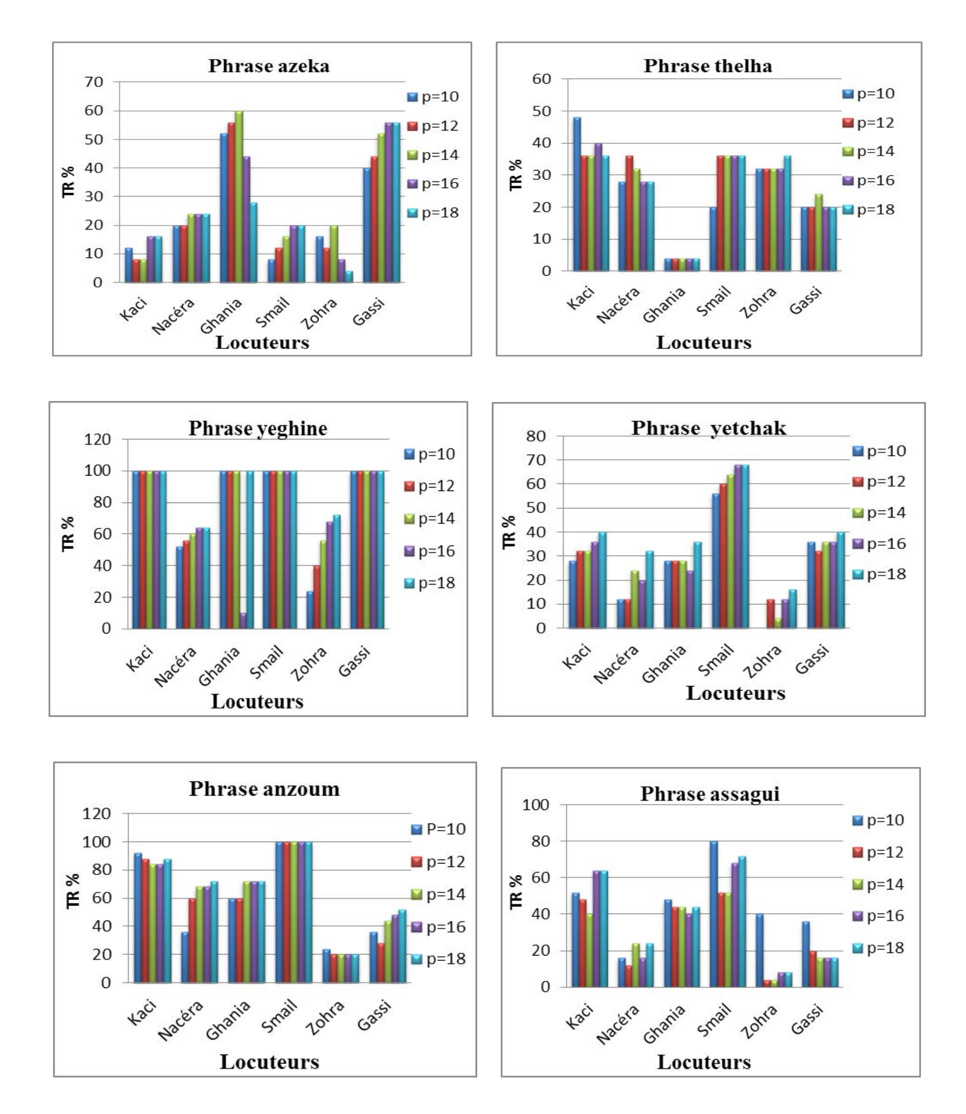


Figure 1 : Influence de l'ordre de prédiction sur le taux de reconnaissance (système dépendant au texte de la langue berbère)

Les ordres de prédiction que nous avons choisi, sont respectivement : 10, 12, 14, 16 et 18 afin d'analyser l'influence de l'ordre de prédiction sur la reconnaissance du locuteur.

D'après les résultats obtenus, les taux de reconnaissance sont supérieurs en mode dépendant de texte comparablement au mode indépendant de texte. De plus, nous remarquons que les taux de bonne classification sont supérieurs pour un ordre de prédiction de l'ordre de 16 et 18 alors qu'ils se dégradent pour un ordre inférieur à 12.

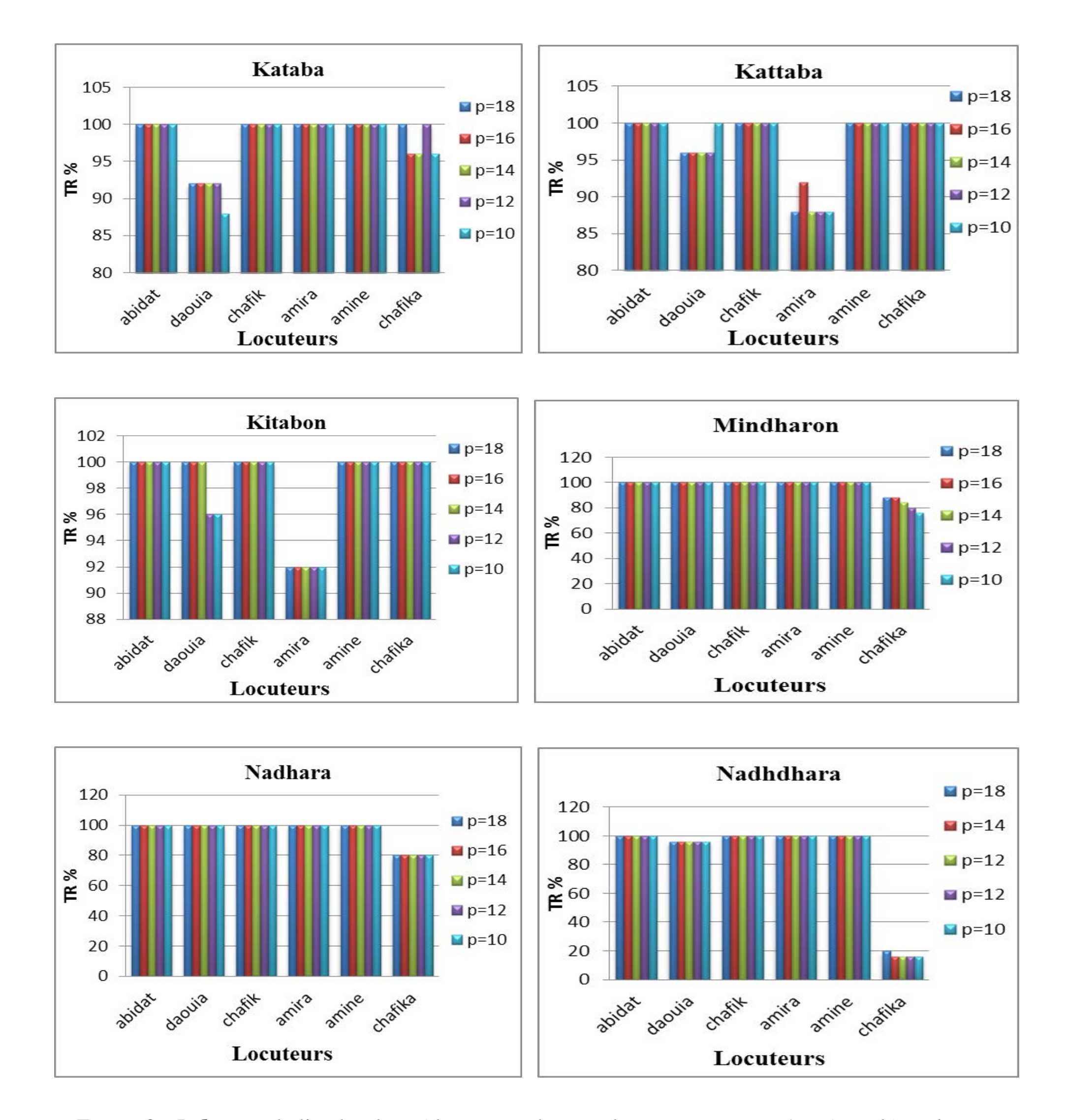


Figure 2 : Influence de l'ordre de prédiction sur le taux de reconnaissance (système dépendant au texte de la langue arabe)

Nous allons par la suite s'intéresser aux dérivées premières et secondes des coefficients LPC, largement utilisés en reconnaissance de la parole et du locuteur, et voir leur influence sur les taux de reconnaissance.

3.3. Dérivées du premier et second ordre

Les dérivées du premier et second ordre sont souvent utilisées en reconnaissance de la parole ou du locuteur. Cela permet d'ajouter de l'information concernant la dynamique du signal. Les coefficients delta (dérivées du premier ordre) sont souvent estimés grâce à l'équation suivante (Holmes, 2003) :

$$\Delta y_t = \frac{\sum_{\tau=1}^{D} (y_{t+\tau} - y_{t-\tau})}{2\sum_{\tau=1}^{D} \tau^2} \qquad (2)$$

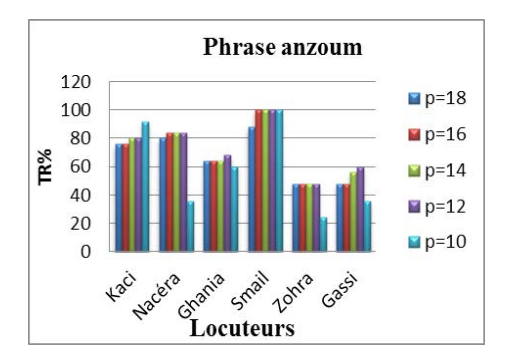
Où D = 1 est le choix habituel pour une fenêtre d'analyse.

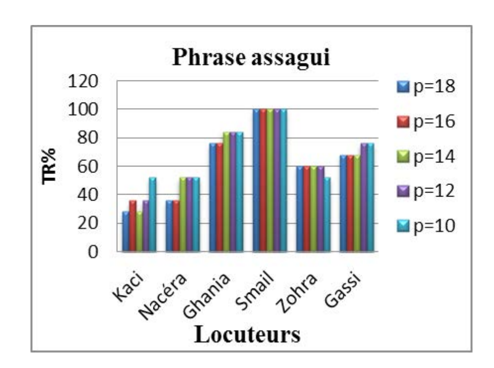
Les coefficients dérivées du second ordre delta-deltas ΔΔ sont estimés de la même manière à partir des coefficients du premier ordre, l'ajout de ces dérivées tend à donner une amélioration supplémentaire.

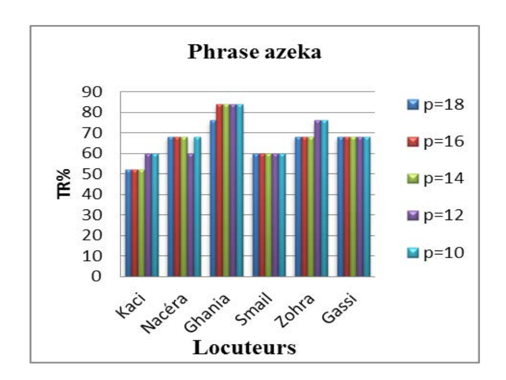
Les histogrammes suivants (Figure 3) permettent une meilleure visualisation des résultats de reconnaissance pour les deux dérivées première et seconde pour les phrases du berbère.

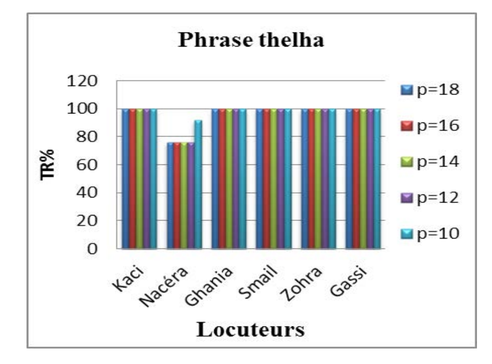
Nous concluons donc que les dérivées premières et secondes améliorent les taux de reconnaissance. En effet, l'ajout des dérivées secondes (qui capturent les changements dans la dynamique du premier ordre) permet une amélioration supplémentaire des taux d'identification mais qui restent petites devant la dérivée première.

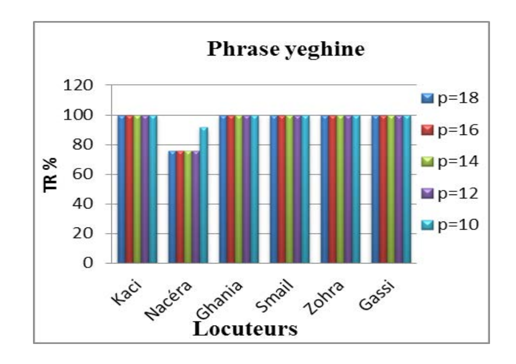
Les avantages de fonctionnalités dérivées sont dus principalement à leur capacité à capturer l'information dynamique. La plupart des systèmes actuels intègrent des fonctionnalités dérivées du premier ordre aux quelles on ajoute une fonction d'énergie, le pitch et beaucoup comprennent les dérivées secondes.











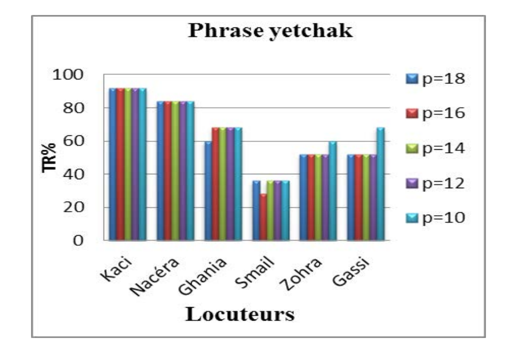


Figure 3 : Taux de reconnaissance pour la première dérivée (système dépendant au texte de la langue berbère).

4. Reconnaissance de locuteurs par fréquence fondamentale

Il existe plusieurs méthodes de détermination de la fréquence fondamentale. Nous présentons dans ce paragraphe les étapes essentielles au calcul du pitch en utilisant le modèle LPC pour rechercher le signal d'excitation.

Le calcul de la fréquence fondamentale se résume comme suit :

Lire l'ensemble des signaux acoustiques d'apprentissage et de test ;

Filtrer le signal par le filtre de Butterworth ;

Fixer la fréquence maximale de recherche à 500 Hz et minimale à 80 Hz ;

Calculer les coefficients LPC sur l'ensemble des trames d'analyse ;

Rechercher le signal d'excitation e(n) tel que :

H(z) = (S(z))/(E(z)) = 1/(A(z)) d'où E(z) = A(z) . S(z) ;

Déduire $e(n) = a(n) * s(n)$;

Calculer la fonction d'auto-corrélation de e(n) ;

Trouver la position du maximum et recherche du 1er pic compris et Tmin et Tmax tel que Tmin=1 /Fmin et Tmax=1/Fmax ;

Affecter la position du 1er pic à la valeur du pitch F0

Fin

La Figure suivante montre la variation du pitch pour la même phrase prononcée par le même locuteur « Gassi » à deux instants différents.

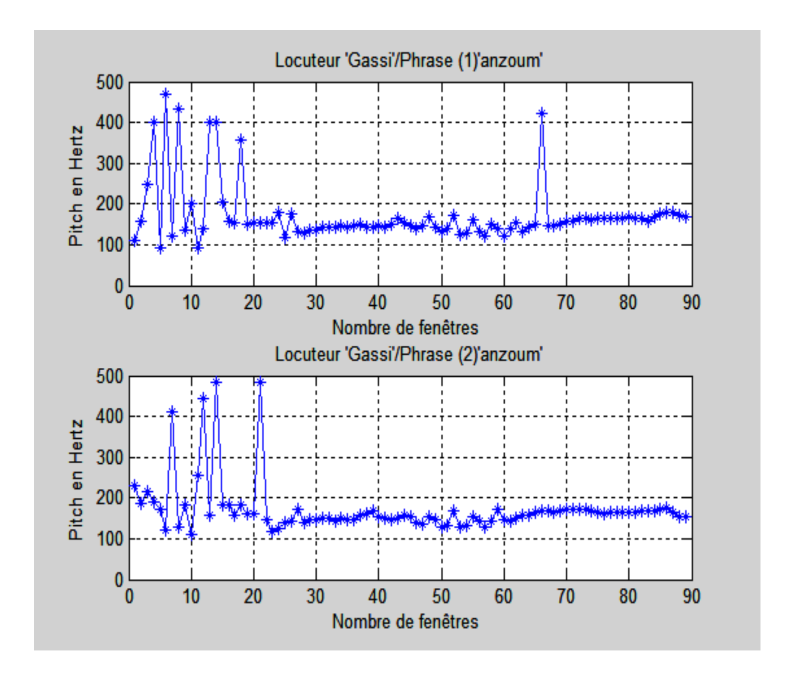


Figure 4 : Locuteur Gassi/Phrase « anzoum ramthane g navdhou » pour deux différentes répétitions

Nous remarquons clairement une légère variation du pitch pour le même locuteur. Ceci démontre la variation intra-locuteurs. Le même phénomène est remarqué lorsque deux locuteurs différents prononcent la même phrase.

La cause principale des différences inter-locuteurs est de nature physiologique. Pour la reconnaissance du locuteur, on cherche à extraire des caractéristiques du signal de parole qui présentent une forte variabilité inter-locuteur (pour pouvoir différencier les locuteurs entre eux) et une faible variabilité intra-locuteur (pour garantir la robustesse du système (Josse, 2003). La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence de base appelée le fondamental. Cette fréquence est différente d'un individu à l'autre et plus généralement d'un genre à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible.

La variabilité inter-locuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux. Cette propriété est valable pour l'arabe standard ou le berbère.

Nous avons présenté les résultats obtenus pour le développement d'un système d'identification basé sur l'analyse LPC et le maximum de vraisemblance. Nous avons aussi amélioré les taux obtenus en ajoutant les dérivées du premier et du second ordre. Les résultats obtenus sont satisfaisants en calculant les dérivées des coefficients LPC.

Par ailleurs, le système basé sur le calcul de la fréquence fondamentale montre des dégradations des systèmes d'identification et ce pour la grande variabilité existante au sein du même locuteur et d'un locuteur à un autre.

5. Conclusion

Nous nous sommes intéressés dans cet article à l'identification de locuteurs en utilisant d'une part les coefficients issus d'une analyse LPC, leurs dérivées premières et secondes et d'autre part la fréquence fondamentale. Le classifieur utilisé est basé sur le maximum de vraisemblance.

Nous avons utilisé deux bases de données acoustiques : la première est celle que nous avons enregistrée au niveau du laboratoire LCPTS, concerne des phrases de la langue berbère, tandis que la deuxième concerne des mots isolés de l'arabe standard déjà existante auparavant.

Nous avons déterminé pour les échantillons d'apprentissage et de test les coefficients LPC, leurs dérivées du premier et second ordre ainsi que la fréquence fondamentale. Par la suite, le classifieur utilisé en choisissant un seuil de 0.75 permet de nous fournir les taux de bonne identification pour chaque locuteur ainsi que le taux global. Des taux de reconnaissance très satisfaisants ont été obtenus (de 80% à 96%) pour la combinaison LPC-dérivée premières et secondes comparablement aux coefficients LPC seuls. Par contre, les taux obtenus pour la fréquence fondamentale sont faibles et ce pour la grande variabilité de ce paramètre lors de l'articulation de signaux acoustiques. Les résultats obtenus sont valables pour les deux bases de données.

Références

Chaker S. (2004). Langue et littérature berbères, Clio.

Chelali F. Z., Djeradi A., Teffahi H. (2012). Caractérisation de voyelles et d'emphatiques berbères en vue d'une identification de locuteurs dépendant de texte. *La 5ème conférence Internationale sur les Technologies d'Information et de Communication pour L'Amazighe*, Rabat, Maroc.

Hasanat M. (2007). Acquisition d'une langue seconde : Les avantages et les entraves de la langue maternelle chez les bilingues français-arabe/arabe-français. *Synergie du monde arabe*, Vol.4, pp. 209-226.

Holmes J., Holmes W. (2003). Introduction to Front-end Analysis for Automatic Speech Recognition, Chapter 10. *Speech Synthesis and Recognition*, Taylor and Francis e-Library.

Mami Y. (2003). Reconnaissance de locuteurs par localisation dans un espace de locuteurs de références. Thèse de doctorat, Ecole nationale supérieure des télécommunications, Paris Telecom, France.

Naït-Zerrad K. (1998). Sur la notation usuelle du berbère – Eléments d'orthographe. Révisé par Chaker S. (2002).

Tanprasert C., Wutiwiwatchai C., Sae-tang S. (2000). Text-dependent speaker identification using neural network on distinctive Thai tone marks. *Technical Journal*, 1(6): 249-253.

Saeed K., Kheir Nammous M. (2007). A Speech-and-speaker identification system: Feature extraction, description, and classification of speech-signal image. *IEEE Transactions on Industrial Electronics*, 54(2): 887-897.

Saidane T., Zrigui M., Ben Ahmed H. (2004). La transcription orthographique-phonétique de la langue arabe. *La 11ème conférence sur le Traitement Automatique des Langues Naturelles, TALN- Récital*, Fès. Maroc.

Toutios A., K. G. Margaritis. (2002). Development of a text-dependent speaker identification system with the OGI toolkit. *Second Hellenic Conference on Al, SETN*, Thessaloniki, Greece, pp. 525-530.

Josse V. (2003). Identification nommée du locuteur: Exploitation conjointe du signal sonore et de sa transcription. Thèse de doctorat, Université du Maine. France.

Détection de l'état émotionnel à partir du signal de parole

Mohamed Réda Moharem   Hocine Teffahi   Leila Falek
L.C.P.T.S, Faculté d'Electronique et d'Informatique
U.S.T.H.B, B.P. 32 El Alia Bab Ezzouar,
Alger 16111, Algérie.
moharem.r@hotmail.fr, {hteffahi,falek.leila}@gmail.com

Résumé

Doter la machine des capacités de compréhension des comportements humains : tel est le défi scientifique autour duquel se rassemblent différentes communautés scientifiques (traitement du signal, traitement automatique du langage, intelligence artificielle, robotique, interaction homme-machine, etc.). L'un des signaux fréquemment utilisé est le signal de la parole, qui est en effet l'une des modalités fondamentales que l'homme utilise pour communiquer. La question n'est plus uniquement de savoir ce qui est dit mais aussi de connaître le contexte de prononciation de la phrase. C'est à ce niveau qu'intervient la dimension émotionnelle. La reconnaissance automatique des émotions est un sujet qui suscite un intérêt croissant dans le domaine du traitement de la parole. Dans cet article, nous allons montrer qu'il est possible d'identifier automatiquement l'état émotionnel en utilisant la prosodie. Le paramètre prosodique utilisé est la fréquence fondamentale de la dernière syllabe du signal de la parole analysé.

1. Introduction

La détection de l'émotion à travers la parole constitue une voie peu explorée. Les méthodes de détection actuelles (Dimitrios et Kotropoulos, 2006) reposent sur des classifieurs qui travaillent sur des indices paralinguistiques (acoustique, prosodie). Hormis quelques marques de disfluences, les indices linguistiques sont rarement considérés. Au mieux, les mots de l'énoncé ne sont considérés que comme marqueurs lexicaux isolés. L'intégration d'indices linguistiques améliore pourtant toujours les performances (Schuller, 2007).

Utiliser uniquement les caractéristiques prosodiques (intensité, pitch et temps de prononciation des voyelles) n'est pas utilisé jusque-là dans l'identification ou la détection de l'émotion, alors que la prosodie constitue une source d'informations importantes car elle est présente dans la stratégie humaine de perception et de compréhension de la parole, ce qui devrait suffire pour la mise au point de ce dernier.

En effet, en moment de colère l'intensité de la voix est élevée, les syllabes sont prononcées plus rapidement et le pitch est plus grave, alors que l'inverse est remarqué pour ces trois caractéristiques quand une personne est émue (Scherer, 1986).

Dans ce travail, les tests ont été effectués sur une base de données en langue allemande, mise au point par 10 Acteurs (5 hommes et 5 femmes), qui prononcent chacun 10 différentes phrases (5 courtes et 5 longues) pouvant être interprétées en plusieurs cas émotionnels. La base de données est constituée de 800 Phrases (Burkhardt *et al*., 2005).

2. Extraction des paramètres prosodiques

L'efficacité de n'importe quel système de détection automatique de l'état émotionnel repose essentiellement sur la sélection du jeu de paramètres pertinents. La prosodie désigne les phénomènes liés à l'évolution dans le temps des paramètres de hauteur, d'intensité et de durée. La valeur de ces paramètres à un instant donné a peu de signification. Ce sont plutôt leurs variations qui peuvent être interprétées. La définition de la prosodie ci-dessus traduit cette préoccupation en faisant intervenir explicitement l'évolution dans le temps des paramètres. Dans ce qui suit, nous donnons une brève description des paramètres utilisés dans ce travail.

2.1. L'énergie

L'énergie est considérée comme le paramètre prosodique le plus simple à calculer. L'estimation de l'énergie d'un signal échantillonné St est définie par :

$$E= \frac{1}{T} \Sigma_{t=1}^{T} S_t^2 \qquad (1)$$

Etant donnée sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{db} = 10 \times \log_{10} (E) \qquad (2)$$

2.2. La durée

La durée est le paramètre le plus difficile à préciser, car rien n'indique comment le système de contrôle, de production ou de perception de parole mesure le temps. Les indices de durée classiques supposent généralement la donnée d'une segmentation, des frontières des unités dont on désire mesurer la durée. La durée d'une unité est alors mesurée par le nombre de trames qui séparent ses frontières de début et de fin. La plupart des systèmes prosodiques utilisent une segmentation basée sur le phonème.

2.3. La fréquence Fondamentale

La mélodie de la voix se traduit sur le plan physique par l'évolution de la fréquence laryngienne (Caractéristique des sons voisés en fonction du temps). La plage de variation moyenne de cette fréquence varie d'un locuteur à l'autre en fonction principalement de son âge et de son sexe et aussi peut enregistrer d'importantes variations chez un même locuteur (Koike *et al*., 1998).

Dans ce travail, nous avons utilisé pour l'extraction de ces paramètres le logiciel PRAAT, un logiciel gratuit et facile à utiliser et à interpréter.

3. Détection de l'état émotionnel

L'émotion est une notion floue et elle est difficilement définissable. Elle présente la particularité d'être idiosyncrasique, c'est-à-dire particulière et propre à chaque individu. De ce fait, plusieurs définitions et rôles ont été donnés à l'émotion.

On peut dire que les manifestations externes des émotions diffèrent d'une personne à une autre selon le caractère, l'âge ainsi que d'autres paramètres, et peuvent être contrôlées. Cependant, les manifestations internes ne peuvent être contrôlées : des hormones sont secrétées, changements dans la respiration et battements de cœurs, et aussi variation des vibrations des cordes vocales.

Donc, il est préférable de se baser sur les paramètres liés avec les manifestations internes (comme le pitch) et non sur ceux liés avec les manifestations externes (comme l'intensité et le débit) (Devillers *et al*., 2010).

3.1. Test de perception

Nous décrivons ici une étude faite avec des étudiants Américains sur la perception des émotions. Les auditeurs ont été divisés en deux groupes et chaque groupe a écouté des énoncés d'une locutrice donnée. Les auditeurs doivent juger s'il y a une émotion dans l'énoncé ou pas (Klasmeyer *et al*., 2000).

Dans un 2ème temps, le même test est effectué mais en écoutant des parties d'énoncés (les énoncés étant divisés en 3 parties : partie initiale, médiane et finale).

Les auditeurs ont passé le test individuellement dans une chambre sourde. Les stimuli étaient présentés aux auditeurs par l'intermédiaire d'un haut-parleur, et cela une fois, dans un ordre aléatoire. L'ordre de la session a été aussi contrebalancé entre les auditeurs. Les auditeurs ont répondu sur le questionnaire pendant deux secondes d'intervalle entre les stimuli d'énoncé et pendant une seconde d'intervalle entre les stimuli de partie. Le test total a duré environ 45 minutes.

Les résultats montrent que l'identification émotionnelle avec les stimuli d'énoncés est plus élevée qu'avec les stimuli de parties. Ce résultat n'est pas surprenant mais n'a pas été démontré de façon expérimentale. Néanmoins, l'indentification de l'émotion pour la partie finale est quasiment équivalente à celle des énoncés.

En conclusion, on obtient de meilleurs résultats pour l'identification de l'état émotionnel en écoutant l'énoncé au complet ou en écoutant la dernière partie de l'énoncé. C'est cette dernière remarque que nous allons exploiter dans ce travail pour détecter automatiquement l'état émotionnel.

Les figures 1, 2, 3 et 4 montrent l'évolution du pitch (ou fréquence fondamentale), pendant l'énoncé de la phrase. Nous remarquons que le pitch ne varie quasiment pas durant la dernière syllabe pour une phrase sans émotion 'Neutre'. Par contre, pour les autres émotions (Joie, Colère et Peur) le pitch varie de façon bien visible. Ceci est conforme aux résultats expérimentaux (Avanzi *et al*., 2011).

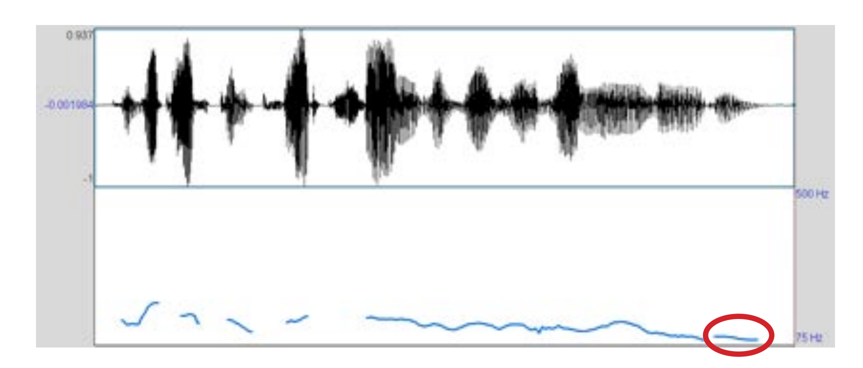


Figure 1 : Variation du Pitch pour l'émotion "Neutre"

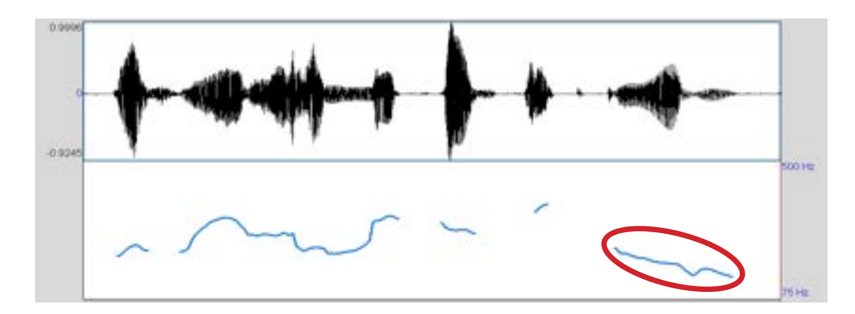


Figure 2 : Variation du Pitch pour l'émotion "Joie"

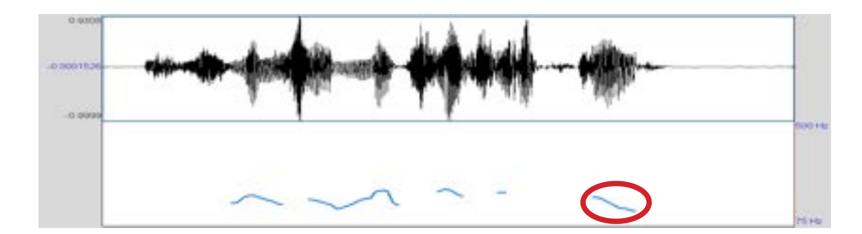


Figure 3 : Variation du Pitch pour l'émotion "Peur"

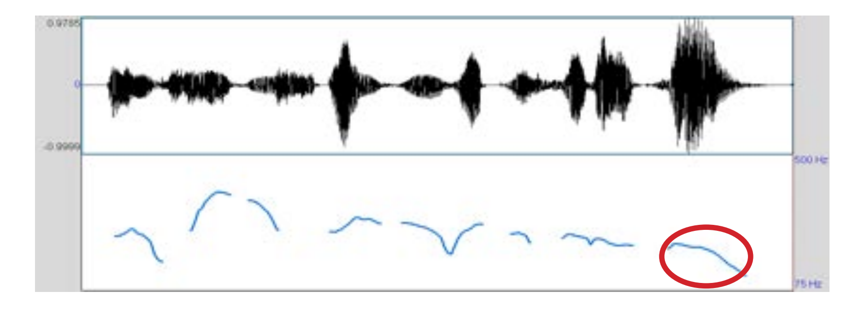


Figure 4 : Variation du Pitch pour l'émotion "Colère"

3.2. Détection automatique de l'état émotionnel

Dans cette partie, un programme est mis au point pour décider si une émotion existe ou pas dans une phrase selon la variation du pitch dans les différentes parties du signal de parole (initiale, médiane et finale). L'identification se fera sur des groupes de phrases contenants la même émotion. Ceci nous permettra de mieux apprécier les résultats d'identification de l'état émotionnel selon l'émotion.

Dans une première étape, il est nécessaire de déterminer le seuil de variation de la fréquence à partir duquel nous pouvons juger qu'il existe une émotion ou pas dans une phrase. Ceci s'obtient en calculant le taux d'identification de l'état émotionnel avec plusieurs intervalles de variations du pitch pour les différentes parties de la phrase.

3.2.1. Détection de l'état émotionnel par analyse de la dernière syllabe de la phrase

Dans cette partie, nous allons voir s'il est possible de détecter l'existence d'une émotion en exploitant uniquement la dernière syllabe de la phrase, en changeant le seuil de variation du pitch. La figure 5 donne les taux de bonne identification des émotions pour différentes plages de variations du pitch durant la dernière syllabe de la phrase.

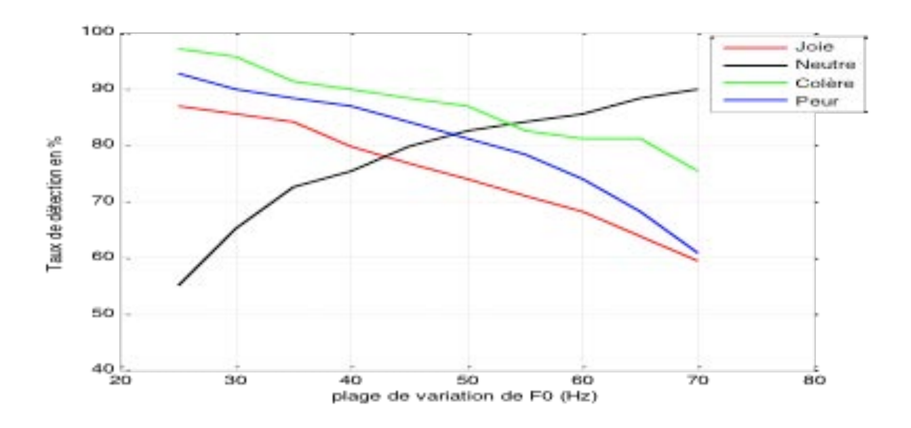
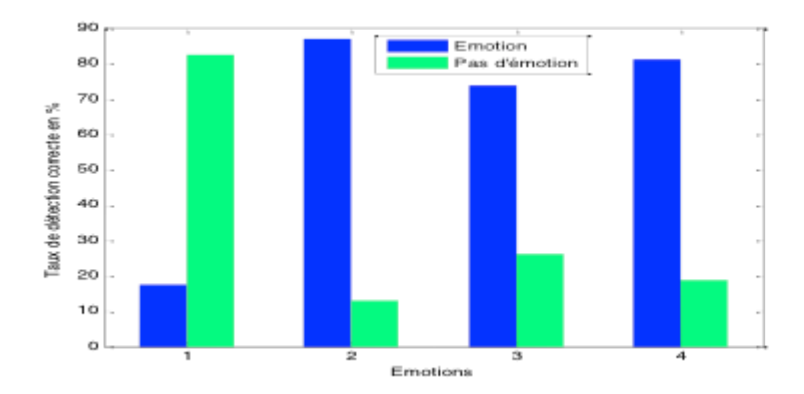


Figure 5 : Taux de reconnaissance de l'état émotionnel pour différentes émotions en utilisant la dernière syllabe de la phrase selon la plage de variation du pitch.

Les taux de bonne identifications entre l'émotion Neutre et les autres émotions se rapprochent le plus lorsque le pitch varie entre 45 et 55 Hz. Nous pouvons donc considérer, en moyenne, qu'une émotion existe si le pitch varie de plus de 50 Hz durant la dernière syllabe de la phrase.

Cette dernière constatation a été appliquée sur toute la base de données et les résultats sont regroupés sur la figure 6. Nous pouvons remarquer clairement que les scores obtenus dépassent dans tous les cas le seuil de 73%. Ce résultat est en parfaite conformité avec les tests de perception.



1 : Neutre, 2 : Colère, 3 : Joie, 4 : Peur

Figure 6 : Taux de reconnaissance de l’état émotionnel pour différentes émotions en utilisant la dernière syllabe de la phrase pour une varation du pitch de 50Hz.

3.2.2. Détection de l’état émotionnel par analyse de la partie médiane de la phrase

Dans cette partie, nous allons refaire le même travail mais en étudiant la variation du pitch dans la partie médiane de la phrase. La figure 7 donne les taux de bonne identification des émotions pour différentes plages de variations du pitch dans la partie médiane de la phrase.

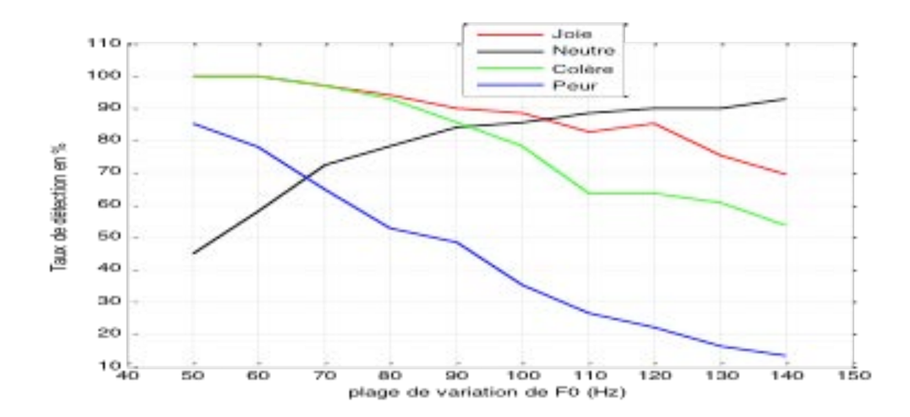
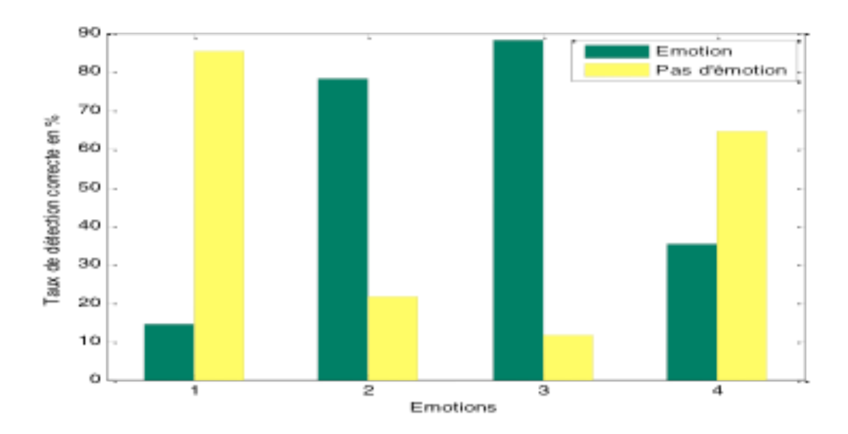


Figure 7 : Taux de reconnaissance de l’état émotionnel pour différentes émotions en utilisant la partie médiane de la phrase selon la plage de variation du pitch.

Les taux de bonne identifications entre l’émotion neutre et les autres émotions se rapprochent le plus lorsque le pitch varie entre 90 et 105 Hz. Nous pouvons donc considérer, en moyenne, qu’une émotion existe si le pitch varie de plus de 100 Hz en analysant la partie centrale de la phrase.

Cette dernière constatation a été appliquée sur toute la base de données et les résultats sont regroupés sur la figure 8. Nous pouvons remarquer clairement que les scores obtenus, pour les cas de la colère et de la joie, est de l'ordre de 80%. Néanmoins, l'émotion Peur est mal identifiée (taux correcte de 35%). Ce résultat nous permet de conclure que la partie médiane de la courbe de variation du pitch est insuffisante pour identifier s'il y a une émotion ou non.



1 : Neutre, 2 : Colère, 3 : Joie, 4 : Peur

Figure 8 : Taux de reconnaissance de l'état émotionnel pour différentes émotions utilisant la partie médiane de la phrase pour une varation du ptich de 100Hz.

3.2.3. Détection de l'état émotionnel par analyse de la première syllabe de la phrase

En ce qui concerne la partie intiale, où la première syllabe prononcée dans une phrase, les résulats des tests sont donnés par la figure 9.

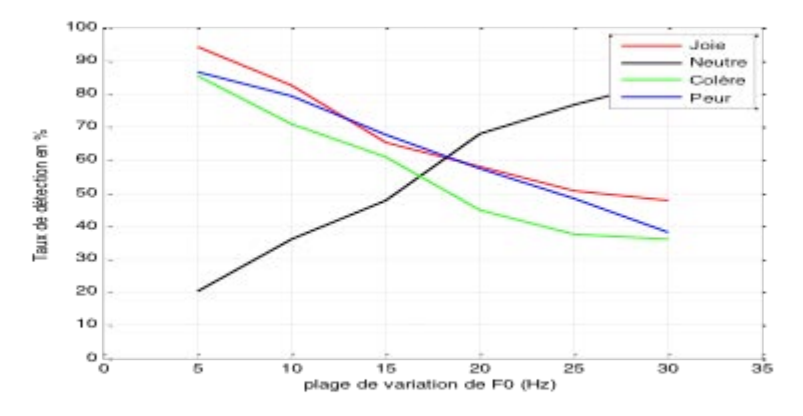
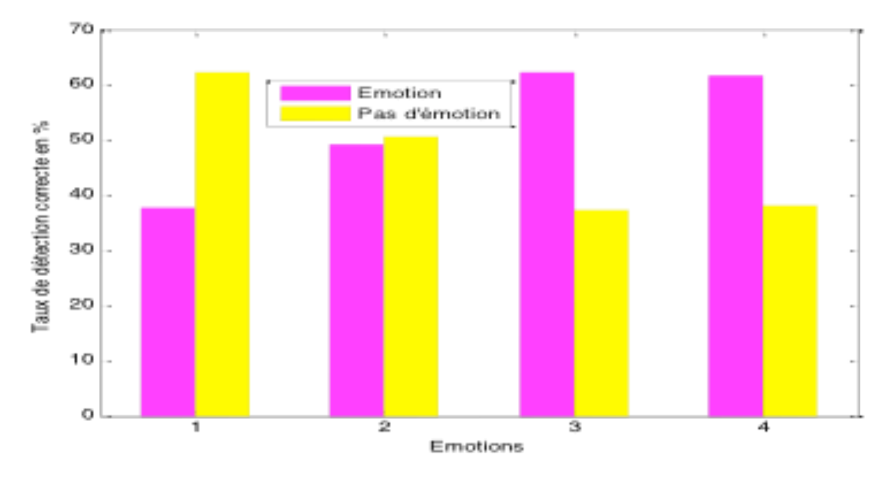


Figure 9 : Taux de reconnaissance de l'état émotionnel pour différentes émotions en analysant la partie initiale de la phrase selon la plage de variation du pitch.

Les taux de bonne identification entre l’émotion Neutre et les autres émotions se rapprochent le plus lorsque le pitch varie entre 15 et 20 Hz. Nous pouvons donc considérer, en moyenne, qu’une émotion existe si le pitch varie de plus de 19 Hz pendant la première syllabe de la phrase.

Cette dernière constatation a été appliquée sur toute la base de données et les résultats sont regroupés sur la figure 10. Nous pouvons remarquer clairement que les scores obtenus dépassent dans les cas de la joie et de la peur le seuil de 60%. Néanmoins, l’émotion Colère est mal identifiée avec ce critère.

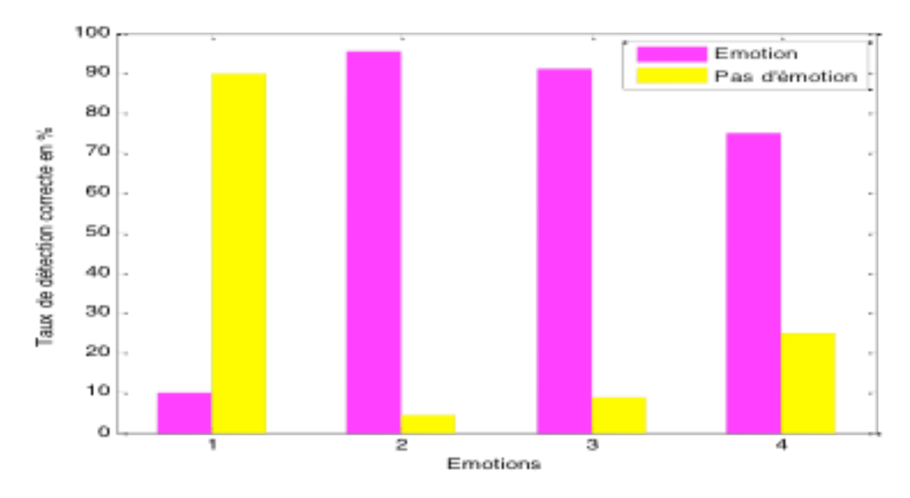


1 : Neutre, 2 : Colère, 3 : Joie, 4 : Peur

Figure 10 : Taux de reconnaissance de l’état émotionnel pour différentes émotions en utilisant la partie initiale de la phrase pour une varation du pitch de 19Hz.

4. Résultats de détection automatique

En utilisant les résultats des différentes expériences utilisées dans les paragraphes précédents, nous avons mis au point une méthode simple et efficace de détection de l’état émotionnel. Nous considérons que la phrase véhicule une émotion si deux critères, parmi les trois cités, au moins sont satisfaits. Les résultats de cette détection automatique sont regroupés sur la figure 11.



1 : Neutre, 2 : Colère, 3 : Joie, 4 : Peur

Figure 11 : Taux de reconnaissance de l'état émotionnel pour les différentes émotions.

D'après ces résultats, nous pouvons affirmer que le contour de la fréquence fondamentale est un très bon critère de détection de l'état émotionnel dans une phrase. Les taux de détection des émotions Colère, Joie et Peur sont assez importants (plus de 75% de bonnes détections). Aussi, le système permet de reconnaître que les phrases neutres ne véhiculent pas d'émotions avec un taux de bonne identification de l'ordre 90%.

Aussi, cette étude nous a permis de vérifier que les émotions Colère, Joie et Neutre, sont mieux identifiées dans la partie médiane et finale de la phrase, alors que pour l'émotion Peur, c'est dans la partie initiale et dans la partie finale que ce taux est plus élevé.

5. Conclusion

Ce travail a permis de vérifier que la détection automatique de l'état émotionnel est possible en utilisant uniquement la mélodie de la phrase. Une étude sur le comportement des gens sous l'effet d'une émotion peut augmenter le taux de reconnaissance pour une obtention de meilleurs résultats, mais ceci reste difficile à cerner vu que ces comportements diffèrent d'une personne à une autre selon le caractère et l'humeur.

A noter que les paramètres prosodiques pour les différentes émotions ne varient pas de la même manière d'un peuple à un autre, en d'autres termes d'une langue à une autre. Ce genre d'étude est donc limité pour une certaine population et ne peut être généralisé sur toutes les classes.

Références

Dimitrios V., Constantine K. (2006). *Emotional speech recognition: Resources, features, and methods*. Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, University Campus, Box 451, Thessaloniki 541 24, Greece.

Schuller R. (2007). The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. *In Proceedings INTERSPEECH 2007, 8th Annual Conference*.

Scherer K. R. (1986). *Vocal affect expression: A review of research paradigms*, Psychological Bulletin, 99, 143-165.

Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. (2005). A database of german emotional speech. *In Proceedings Interspeech 2005*, Lissabon, Portugal, 1517–1520.

Koike K., Suzuki H., H Saito. (1998). Prosodic parameters in emotional speech, *5th Int. Conf. on Spoken Language Processing (ICSLP 98),* Sydney, Australia, November 30 – December 4, paper 0996.

Devillers L., Vidrascu L., Layachi O. (2010). Automatic detection of emotion from vocal expression. *A blueprint for an affectively competent agent, Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. *K. Scherer, T. Bänziger and E. Roach*, Oxford University Press: 232-244.

Klasmeyer G. (2000). An automatic description tool for time contours and long-term average voice features in large emotional speech databases, *ITRW on speech and emotion*, Newcastle, Northern Ireland, UK, September 5-7, 66-71.

Avanzi M., Obin N., Lacheret-Dujour A., Victorri B. (2011). Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree. *InterSpeech 2011*. Florence: 2033-2036.

TICAM'14, p. 169-176

Evaluation de Quelques Méthodes de Segmentation "Application aux Caractères Tifinaghes"

Abdelkrim Maarir **Ilhame Agnaou** **Belaid Bouikhalene**
Laboratoire Traitement de l'Information et Aide à la Décision,
Université Sultan Moulay Slimane - F.S.T Beni Mellal, Maroc
a.maarir@usms.ma

Résumé

Le traitement d'image est l'un des domaines de recherche les plus avancés, l'image occupe une place particulière dans de nombreux domaines de la vie quotidienne (manuscrit, les images satellites, la publicité, l'industrie, etc.). Dans cet article, nous allons mettre l'accent sur les images des caractères tifinaghes, en présentant une comparaison de quelques méthodes de segmentation (détection) existantes classées selon leurs catégories et réaliser une étude comparative sur la base de la qualité de la segmentation et le temps de calcul.

1. Introduction

La segmentation joue un rôle prépondérant dans le traitement d'image. Elle est réalisée avant les étapes d'analyse et de prise de décision dans plusieurs processus d'analyse d'image, telle que la détection des objets. Elle facilite la localisation et la délimitation des entités présentes dans l'image.

La segmentation d'image est une opération de traitement d'image qui a pour but de rassembler des pixels entre eux suivant des critères prédéfinis. Les pixels sont ainsi regroupés en régions, qui constituent une partition de l'image. Il peut s'agir par exemple de séparer les objets du fond si le nombre de classes est égal à deux. Elle est appelée dans ce cas binarisation.

L'objectif de cet article est de présenter quelques méthodes de segmentation, telles que la segmentation par classification, la croissance de régions, la détection de contours - régions et la segmentation par Quadtree.

Cet article est organisé comme suit : Section 2 est consacrée à la base de données des caractères tifinaghes manuscrits. Dans la section 3, nous décrivons les différentes méthodes de segmentation d'images. Les résultats expérimentaux et l'application aux caractères tifinaghes sont présentés à la section 4. La section 5 conclut cet article.

2. Base de données des caractères tifinaghes

Le Tifinaghe est un alphabet utilisé par les amazighes, essentiellement les Touaregs. Il était un alphabet consonantique, puis il a subi des modifications et des variations inévitables

depuis son origine jusqu'à nos jours.

Depuis l'année 2003 Tifinaghe-IRCAM est devenu le système graphique officiel pour l'écriture de la langue amazighe au Maroc. La figure 1 montre les 33 caractères tifinaghes.

Figure 1 : Tifinaghe - IRCAM

3. Les Méthodes de segmentation

3.1. Segmentation par classification

Elle permet d'effectuer une classification dans l'espace de luminance, puis à extraire une partition par la méthode seuillage en utilisant les niveaux de gris présents dans l'image (MacQueen, 1967), comme par exemple la moyenne des niveaux de gris de l'ensemble des pixels, ou la médiane, permettant de construire n classes d'intensité (Tatiraju et Mehta, 2008). La figure 2 présente un exemple d'une image qui est composée de fond et d'un objet.

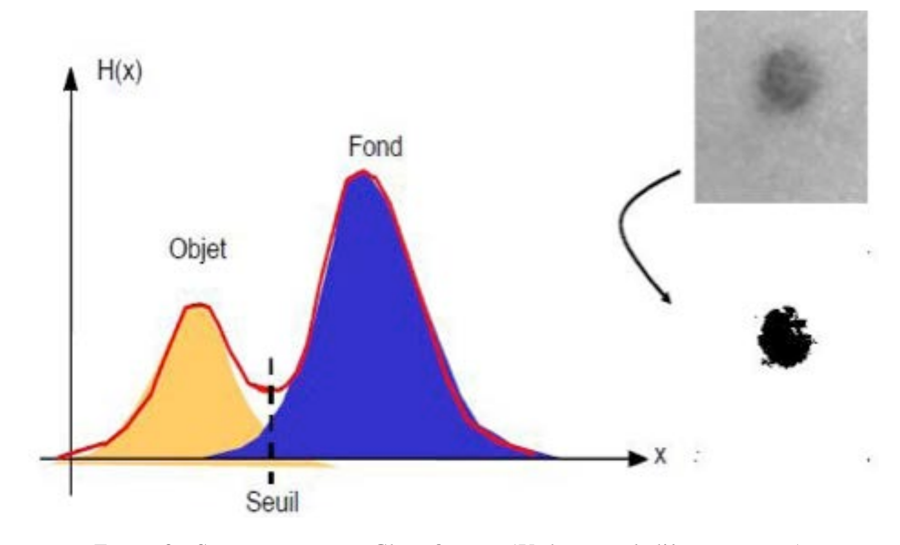


Figure 2 : Segmentation par Classification (Utilisation de l'histogramme)

Le seuillage a pour objectif de segmenter une image en plusieurs classes en n'utilisant que l'histogramme (MacQueen, 1967).

3.2. Segmentation par croissance de régions

Cette segmentation permet de regrouper itérativement l'ensemble connexe de pixels jusqu'à l'obtention de régions maximales selon un critère d'homogénéité. C'est une méthode locale récursive qui a pour principe de faire croître une région avant de passer à la suivante.

L'image est segmentée en régions. Les pixels adjacents similaires selon un certain critère d'homogénéité sont regroupés en régions distinctes. Il y a croissance suivant un critère de similarité jusqu'à atteindre le critère d'arrêt : par exemple convexité maximum, etc.

La croissance par région et la division-fusion sont des techniques courantes de cette catégorie contrairement à l'extraction des contours qui s'intéresse aux bords des régions. La segmentation en régions homogènes vise à segmenter l'image en se basant sur des propriétés intrinsèques des régions (voir Figure 3).

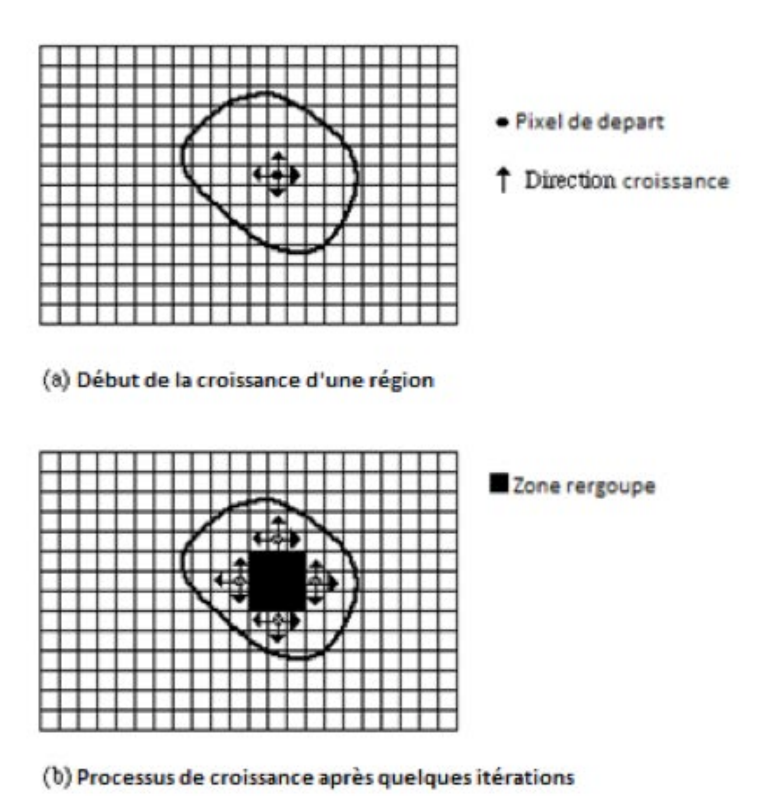


Figure 3 : Segmentation par croissance de régions

3.3. Segmentation par détection de contours et de régions

La segmentation par contours permet de détecter les limites entre les régions de l'image. Les détecteurs de contours utilisés peuvent être simples, comme les opérateurs de Sobel ou de Roberts, ou plus complexes tel que l'opérateur de Canny [1].

Généralement, plusieurs zones de différents niveaux de gris correspondent aux différents objets de la scène. Ainsi, il faut trouver une manière de repérer ces objets dans l'image. Deux grandes approches peuvent être envisagées pour extraire les zones pertinentes de l'image :

L'approche région qui permet de constituer une partition I de l'image en regroupant tout les régions homogènes X_1, X_2, ..., X_n, tel que :

1- I = Union X_i, i=1,..., n ;

2- Pour tout i, X_i est connexe ;

3- Pour tout i, H [X_i] est vrai (critère H) ;

4- Pour tout couple (X_i, X_j) de régions voisines, H[X_i, X_j] est faux.

L'approche contour qui permet de rechercher les discontinuités dans une scène (Maitre, 2005). La détection de contours permet de repérer dans les images les objets qui s'y trouvent avant d'appliquer le traitement uniquement sur ces objets. Les contours dans une image proviennent des discontinuités de la fonction de préfectance (texture, ombre) et des discontinuités de profondeur (bords de l'objet).

Les régions sont délimitées par les contours des objets (séparation) qu'elles représentent (voir figure 4).

Figure 4 : Détection de contours

1 http://en.wikipedia.org/wiki/Canny_edge_detector

4. Segmentation par Quadtree

Un quadtree est une structure de données de type arbre dans laquelle chaque nœud interne a exactement quatre sous nœuds[2]. La région quadtree représente une partition de l'espace en deux dimensions par la décomposition de la région en quatre quadrants égaux et à chaque itération la segmentation crée des objets carrés réguliers, où la taille est définie par la variation des objets.

Les objets sont créés en divisant des objets plus grands jusqu'à ce que les objets qui en résultent soient à la limite supérieurs de la variation admise comme le montre la Figure 5.

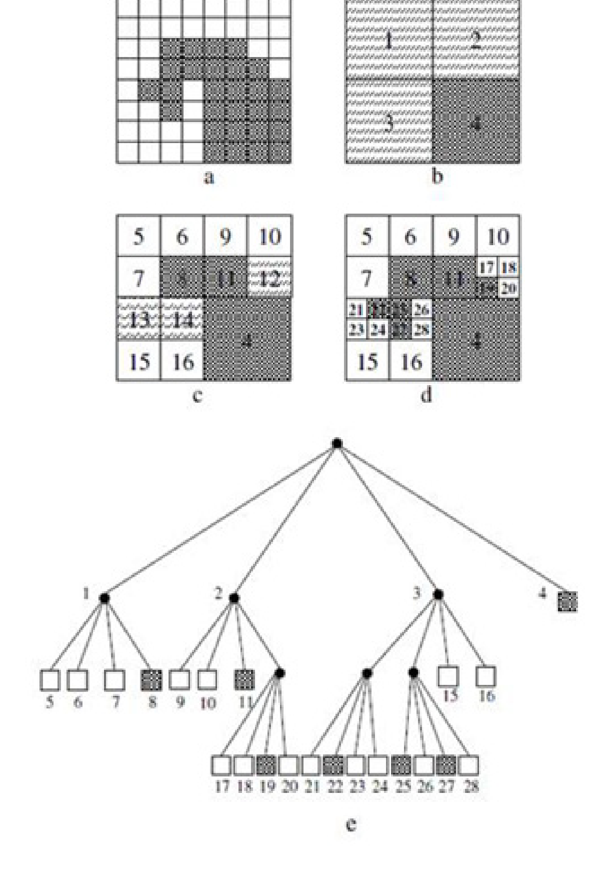
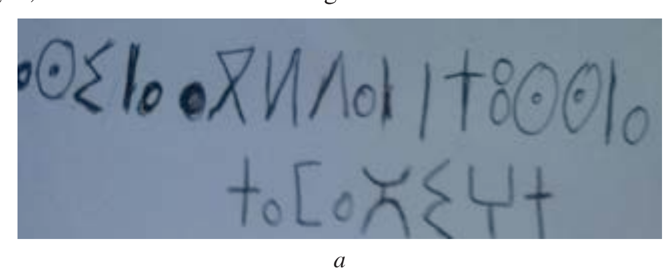


Figure 5 : a) Image original, b) 1[er] niveau de séparation, c) 2[ème] niveau, d) 3[ème] et final niveau, e) Quad-tree correspondant.

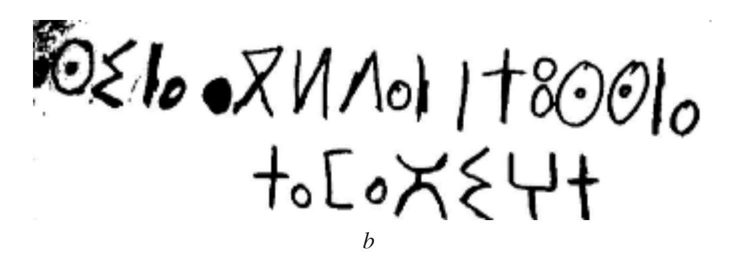
2 http://en.wikipedia.org/wiki/Quadtree

5. Résultats expérimentaux

Dans cette section, nous avons évalué les méthodes de segmentation mentionnées précédemment, nous montrons les résultats des expérimentations appliquées sur une image contenant du texte en tifinaghe. Les Figures 6 et 7 ci-dessous présentent un exemple du texte écrit en tifinaghe, sa binarisation et son histogramme horizontal.



a



b

Figure 6: Image du texte amazighe (a) et sa binarisation (b)

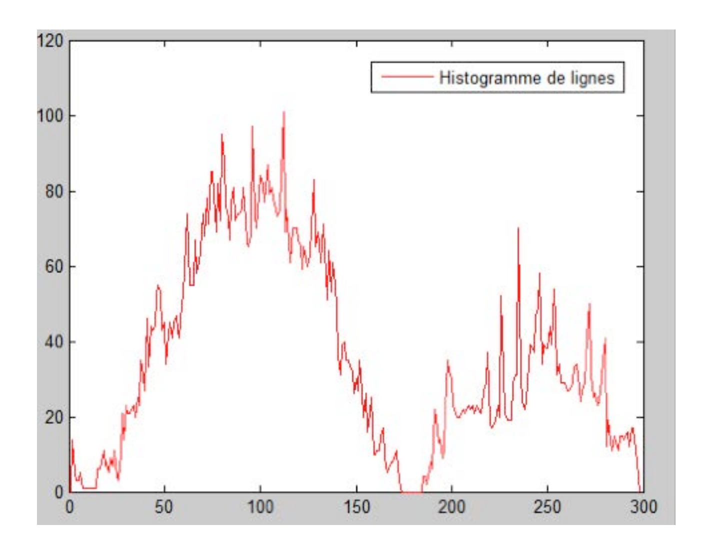
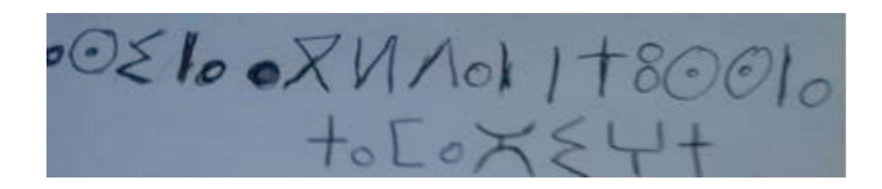


Figure 7: Histogramme horizontal

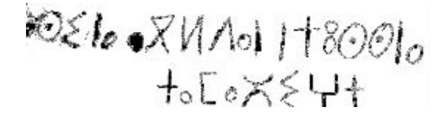
Cette étude est faite sur l'image de la Figure 6, dont les caractéristiques sont présentées dans le Tableau ci-dessous :

Caractéristique	Valeur
Mean	222.404
Ecart	84.0309
Max value	178
Min value	3

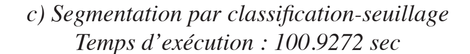
Tableau 1 : Les caractéristiques de l'image de la Figure 6



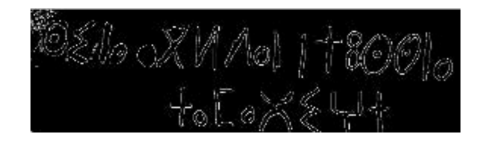
a) Image Original



b) Segmentation par croissance de régions
Temps d'exécution : 143.8361 sec



c) Segmentation par classification-seuillage
Temps d'exécution : 100.9272 sec



d) Détection de contour par canny
Temps d'exécution : 0.3932 sec

e) Segmentation par Quadtree
Temps d'exécution : 5.099 sec

Figure 8: Application des méthodes de segmentation sur une image contenant du texte en tifinaghe

La figure 8 montre les résultats de l'application des différentes méthodes de segmentation sur le texte amazighe. On peut dire que la segmentation contour et quadtree (Figure 8d-e) sont les méthodes qui permet de détecter les contours avec le plus de détails. Les contours sont ici beaucoup plus nets, par contre la segmentation par croissance de régions et la segmentation

par classification (Figure 8b-c) montrent que la détection des caractères et moins performante et le taux d'exécution est très élevé.

6. Conclusion

Cet article présente une comparaison entre deux grandes approches de segmentation : approche région et approche contour en se basant sur quelques méthodes parmi les plus connues, à savoir la segmentation par classification, la croissance de régions, la détection de contours-régions et la segmentation par Quadtree. Leur application sur les caractères tifinaghes, pour l'extraction des différentes zones homogènes distinctes d'une image dont les membres partagent une diversité de propriétés communes, montre que dans le cas où la qualité de la segmentation et le taux d'exécution sont favorisés l'utilisation de la segmentation par contours est souhaitable.

Références

MacQueen J. (1967). Some methods of classification and analysis of multivariate observations. *Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, Etat-Unis.

Maitre H. (2005). « La détection des contours dans les images ». Chapitre 11 du Polycopié de cours : Le traitement des images, tome 2, Département TSI-Télécom-Paris.

Tatiraju S., Mehta A. (2008). Image segmentation using k-means clustering, EM and Normalized cuts. *Proceeding of Symposium of Discrete Algorithm*, juillet 2008.

TICAM'14, p. 177-187

Code Braille et la Reconnaissance d'un Document Ecrit en Tifinaghe

Rachid El Ayachi Mohammed Fakir
Mustapha Oujaoura Brahim Minaoui
Laboratoire de Traitement de l'Information et Aide à la Décision (TIAD)
Département d'Informatique, Facultés des Sciences et Techniques
Université Sultan Moulay Slimane, Béni Mellal, Maroc
rachid.elayachi@usms.ma, M.Mustapha.Oujaoura@ieee.org, {fakfad,bra_min}@yahoo.fr

Résumé

Nous avons proposé dans ce travail un système de reconnaissance des documents en Tifinaghe, en affichant le code Braille correspondant qui s'articule essentiellement autour de deux parties. Une partie d'apprentissage et une de reconnaissance. Dans la première partie nous avons effectué l'acquisition des données, la définition des attributs caractérisant les données (les paramètres caractéristique vus comme les plus pertinents : les moments Zernike et les moments de Legendre), et la construction de la base d'apprentissage. A la fin de cette phase nous obtenons une matrice qui contient les données et les attributs les caractérisant. Dans la deuxième partie, nous avons utilisé les réseaux de neurones multicouches. Nous avons fournis à notre système une image de document en Tifinaghe. La première étape consiste à binariser, normaliser et segmenter l'image d'entrée en caractères, la seconde étape consiste à effectuer un certain nombre de mesures qui vont permettre de caractériser la forme que nous transformerons ensuite en un vecteur. Dans la troisième étape, ce vecteur est présenté à l'entré d'un classifieur, qui nous fournira en sortie la classe du caractère. Le taux de reconnaissance obtenu est environ de 80 %, qui est très satisfaisant, mais comme dans tout logiciel, des améliorations restent toujours à apporter.

1. Introduction

Le système Braille est un système adopté pour aider les aveugles et les malvoyants à s'intégrer dans les différents secteurs de la vie. Ce système est inventé par Louis Braille à l'âge de 15 ans. Il se base sur un ensemble de points (qui sont au nombre de 6) disposés dans un tableau (Figure1) de deux colonnes, chaque colonne comporte 3 points. L'absence ou la présence des points représentent les différentes combinaisons possibles (26 combinaisons) et chaque combinaison peut être attribuée à un caractère.

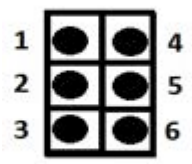


Figure1 : Tableau de Braille

Alors, concernant les caractères Tifinaghes préconisés par l'IRCAM, qui sont au nombre de 33, ils peuvent être exprimés à l'aide du code Braille (Tableau 1). Ce tableau est déduit à partir de la correspondance entre les caractères Tifinaghes, Arabe et Latin ainsi que la proposition des codes Brailles pour le reste (les caractères Tifinaghes qui n'ont pas de correspondance).

Numéro du caractère	Caractère	Code Braille
1	ⴰ	⠁ (1)
2	ⴱ	⠃ (1-2)
3	ⴳ	⠛ (1-2-4-5)
4	ⴳⵯ	⠻ (1-2-4-5-6)
5	ⴷ	⠙ (1-4-5)
6	ⴹ	⠹ (1-4-5-6)
7	ⴻ	⠑ (1-5)
8	ⴼ	⠋ (1-2-4)
9	ⴽ	⠅ (1-3)
10	ⴽⵯ	⠿ (1-2-3-4-5-6)
11	ⵀ	⠓ (1-2-5)
12	ⵃ	⠳ (1-2-5-6)
13	ⵄ	⠕ (1-3-5)
14	ⵅ	⠭ (1-3-4-6)
15	ⵇ	⠟ (1-2-3-4-5)
16	ⵉ	⠊ (2-4)
17	ⵊ	⠚ (2-4-5)
18	ⵍ	⠇ (1-2-3)
19	ⵎ	⠍ (1-3-4)
20	ⵏ	⠝ (1-3-4-5)
21	ⵓ	⠥ (1-3-6)
22	ⵔ	⠗ (1-2-3-5)
23	ⵕ	⠷ (1-2-3-5-6)
24	ⵖ	⠪ (2-4-6)
25	ⵙ	⠎ (2-3-4)
26	ⵚ	⠮ (2-3-4-5-6)
27	ⵛ	⠉ (1-4)
28	ⵜ	⠞ (2-3-4-5)
29	ⵟ	⠏ (1-2-3-4)
30	ⵡ	⠺ (2-4-5-6)
31	ⵢ	⠽ (1-3-4-5-6)
32	ⵣ	⠵ (1-3-5-6)
33	ⵥ	⠶ (2-3-5-6)

Tableau1 : Correspondance des caractères Tifinagh en code Braille

Le code braille se caractérise par l'écriture dans le recto de la page (la création des creux) et la lecture dans le verso de la page (l'utilisation des doits pour parcourir les bosses formés). Par exemple, pour lire le mot « T I F I N A G H E » dans ce sens (verso de la page), l'écriture du mot doit être effectuée dans le sens inverse de la façon suivante (recto de la page) : « E H G A N I F I T ». Alors, dans le cas d'un texte Tifinaghe écrit en code Braille, la lecture s'effectue de la gauche vers la droite, par contre l'écriture s'effectue dans le sens inverse (Figure2).

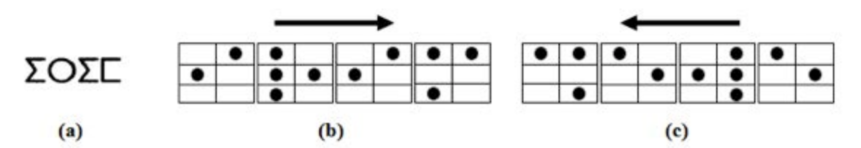


Figure2 : (a) Texte en Tifinaghe, (b) Lecture en Braille et (c) Ecriture en Braille

La structure de l'article est organisée de la façon suivante : Section 2 s'intéresse à la présentation du système de reconnaissance élaboré. Section 3 représente la phase d'extraction d'attributs détaillant les différentes méthodes utilisées. Section 4 traite la phase de la classification. Section 5 illustre les résultats obtenus et la section 6 donne les différentes conclusions tirées à partir de ce travail.

2. Système de reconnaissance

Le domaine de la reconnaissance des caractères a connu une évolution dernièrement, surtout pour les caractères Tifinaghes. Cette évolution est prouvée par le nombre d'articles publiés. Parmi ces publications, nous citons :

- Amrouch *et al*. (2010) ont présenté un système automatique de reconnaissance de caractères amazighes basé sur le Modèle de Markov Caché (HMM).
- El Ayachi *et al*. (2011) ont proposé un système de reconnaissance des caractères Tifinaghes utilisant les réseaux de neurones multicouches. Dans la phase d'extraction des paramètres, la transformation de Walsh est utilisée.
- Es-Saady *et al*. (2011) ont proposé un système de reconnaissance des caractères amazighes. Une méthode basée sur les axes horizontal et vertical est utilisée pour le calcul des attributs. Les réseaux de neurones sont utilisés dans la phase de classification.
- El Ayachi *et al*. (2012) ont développé une méthode hors-ligne où les moments invariants sont calculés à l'aide de la transformation de Fourier.
- Oujaoura *et al*. (2013) ont élaboré un système de reconnaissance qui se base sur la combinaison de plusieurs approches d'extraction et de classification.
- …etc

Le système élaboré comporte plusieurs phases :

- **Prétraitement** : après l'acquisition de l'image d'entrée, un ensemble de fonctions s'appliquent afin d'obtenir une version d'image nettoyée qui va être utilisée

efficacement dans les prochaines étapes. La binarisation est la première fonction à appliquer pour transformer l'image d'entrée en binaire. La normalisation est la deuxième fonction qui a comme pour objectif d'éliminer les zones inutiles dans la raison de réduire le temps de calcul et de minimiser l'espace mémoire utilisé. Finalement, la segmentation, à l'aide des histogrammes horizontal et vertical, est adoptée pour segmenter le texte en lignes, ensuite les lignes en caractères.

- **Extraction** : permet de convertir l'image en un vecteur contenant les caractéristiques de l'image. Les méthodes utilisées dans ce stade sont : les moments de Legendre et les moments de Zernike.
- **Classification** : la robustesse du système se base sur les résultats fournis dans cette étape, c'est pour cela, il faut adopter les approches les plus fiables dans ce domaine, à savoir les réseaux de neurones.

3. Extraction

L'extraction d'attributs est une phase importante dans le système de reconnaissance qui permet de réduire les informations de l'image en un vecteur de paramètres. Ce vecteur sera utilisé ultérieurement dans les étapes qui suivent.

Parmi les approches qui peuvent être adoptées dans cette phase, on trouve les descripteurs invariants (El Ayachi, 2012) (moments de Legendre, moments de Zernike, …) qui prennent en considération la représentation de l'image en quelques situations : la translation, la rotation et l'échelle.

3.1 Moments de Legendre

Les moments de Legendre ont été introduits par Teague (Alt, 1962). Ils ont été utilisés dans plusieurs applications de reconnaissance de formes (Teague, 1980). La propriété d'orthogonale pour les polynômes de Legendre n'implique pas la redondance ou le chevauchement d'informations entre les moments d'ordres différents. Cette propriété permet à la contribution de chaque instant d'être unique et indépendante de l'information dans une image (Chonga *et al*., 2004).

Pour une image discrète de taille *M x N* et d'intensité *I(x,y)*, les moments de Legendre se calculent de la façon suivante :

$$L_{pq} = \lambda_{pq} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} P_p(x_i)\, P_q(y_j)\, I(x,y) \qquad (1)$$

Où $\lambda_{pq} = \frac{(2p+1)(2q+1)}{M \times N}$, x_i et y_j désignent les coordonnées normalisées d'un pixel de la gamme [-1,+1] avec :

$$\begin{cases} x_i = \dfrac{2x-(M-1)}{M-1} \\ y_j = \dfrac{2y-(N-1)}{N-1} \end{cases} \qquad (2)$$

$P_p(x)$ est le polynôme de Legendre d'ordre p défini par :

$$P_p(x)=\sum_{k=0}^{p}\left\{\frac{(-1)^{\frac{p-k}{2}}(p+k)!\,x^k}{2^p k!\left(\frac{p-k}{2}\right)!\left(\frac{p+k}{2}\right)!}\right\}_{p-k=even} \quad (3)$$

Pour accélérer les calculs, la formule récurrente suivante peut être utilisée :

$$\begin{cases} P_p(x)=\dfrac{(2p-1)x}{p}P_{p-1}(x)-\dfrac{(p-1)}{p}P_{p-2}(x) \\ P_1(x)=x, \quad P_0(x)=1 \end{cases} \quad (4)$$

3.2 Moments de Zernike

Les moments de Zernike sont construites en utilisant un ensemble de polynômes complexes qui forment un ensemble orthogonal complet sur le disque de l'unité avec $(x^2 + y^2) = 1$:

$$Z_{mn}=\frac{m+1}{\pi}\int_x\int_y I(x,y)\left[V_{mn}(x,y)\right]dxdy$$

Où *m* et *n* définissent l'ordre du moment et le niveau du gris d'un pixel de l'image $I(x, y)$.

Les polynômes de Zernike $V_{mn}(x,y)$ sont exprimés en coordonnées polaires de la façon suivante:

$$V_{mn}(r,\theta)=R_{mn}(r)e^{-jn\theta} \quad (6)$$

Où $R_{mn}(r)$ est le polynôme orthogonal radial :

$$R_{mn}(r)=\sum_{s=0}^{\frac{m-|n|}{2}}\frac{(-1)^s r^{m-2s}(m-s)!}{s!\left(\frac{m+|n|}{2}-s\right)!\left(\frac{m-|n|}{2}-s\right)!}$$

4. Classification

Les résultats trouvés dans la phase de la classification déterminent l'efficacité d'un système de reconnaissance. C'est pour cela, il faut utiliser des approches robustes dans le domaine de la classification qui augmentent le taux de reconnaissance et diminuent le taux d'erreurs. Parmi ces approches, on trouve les réseaux de neurones (El Ayachi *et al*., 2011; El Ayachi *et al*., 2012; Oujaoura *et al*., 2013; Oujaoura *et al*., 2014).

Un réseau de neurones est composé d'éléments simples (ou neurones) fonctionnant en parallèle. Ces éléments ont été fortement inspirés du système nerveux biologique. Il est constitué d'un graphe pondéré et orienté dont les nœuds symbolisent les neurones formels.

Ces derniers possèdent une fonction d'activation qui permet d'influencer les autres neurones du réseau. Les connexions entre les neurones, que l'on nomme liens synaptiques, propagent l'activité des neurones avec une pondération caractéristique de la connexion. On peut entraîner un réseau de neurones pour une tâche spécifique en ajustant les valeurs des connections (ou poids) entre les éléments (neurones).

La Figure 3 représente l'exemple de réseau de neurones utilisé, c'est un réseau multicouche contenant trois couches : une couche d'entrée de n neurones, une couche de sortie de m neurones et une couche caché de r neurones.

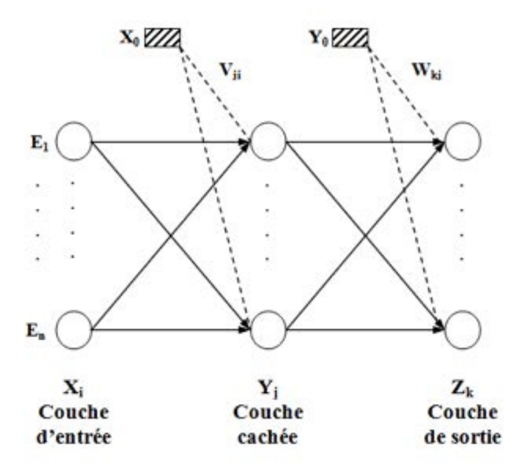


Figure3: Réseaux de neurones

Cet exemple de réseaux (Figure 3) comporte :

- Une couche d'entrée à n cellules d'entrées (n éléments du vecteur d'entrée) $E_i = X_i$;
- Une couche cachée de m neurones d'activations Y_j ;
- Une couche de sortie de r neurones d'activations Z_k ;
- $n \times m$ connexions entre la couche d'entrée et la couche cachée, chacune pondérée par V_{ji} ;
- $m \times r$ connexions entre la couche cachée et la couche de sortie, chacune pondérée par W_{kj} ;
- V_{ji} et W_{kj} sont des poids synaptiques ;
- X_0 et Y_0 sont des seuils.

La fonction d'activation utilisée est la fonction sigmoïde définie de la façon suivante :

$$f(a) = \frac{1}{1 + \exp(-a)} \qquad (8)$$

La propagation vers la couche cachée se fait à l'aide de la formule suivante :

$$Y_j = f\left(\sum_{i=1}^{n} X_i V_{ji} + X_0\right) = \frac{1}{1+\exp\left(-\left(\sum_{i=1}^{n} X_i V_{ji} + X_0\right)\right)} \qquad (9)$$

La propagation de la couche cachée vers la couche de sortie effectue comme suit :

$$Z_k = f\left(\sum_{j=1}^{m} Y_j W_{kj} + Y_0\right) = \frac{1}{1+\exp\left(-\left(\sum_{j=1}^{m} Y_j W_{kj} + Y_0\right)\right)} \qquad (10)$$

Au niveau de la couche de sortie, l'erreur entre la sortie désirée S_k et la sortie réelle Z_k est calculée par :

$$E_k = Z_k(1 - Z_k)(S_k - Z_k) \qquad (11)$$

L'erreur calculée est propagée sur la couche cachée en utilisant la formule suivante :

$$F_j = Y_j(1 - Y_j)\sum_{k=1}^{r} W_{kj}.E_k \qquad (12)$$

La correction des poids de connexions entre la couche d'entrée et la couche cachée utilise les formules suivantes :

$$\Delta V_{ji} = \eta.X_i.F_j \quad \text{et} \quad \Delta Y_0 = \eta.F_j \qquad (13)$$

La modification des connexions entre la couche cachée et la couche de sortie se calcule comme suit :

$$\Delta W_{kj} = \eta.Y_j.E_k \quad \text{et} \quad \Delta X_0 = \eta.E_k \qquad (14)$$

Avec η une constante positive appelée taux d'apprentissage (un paramètre à déterminer empiriquement).

5. Résultats

Les Figures 4, 5 et 6 représentent des exemples de résultats obtenus par le système de reconnaissance proposé.

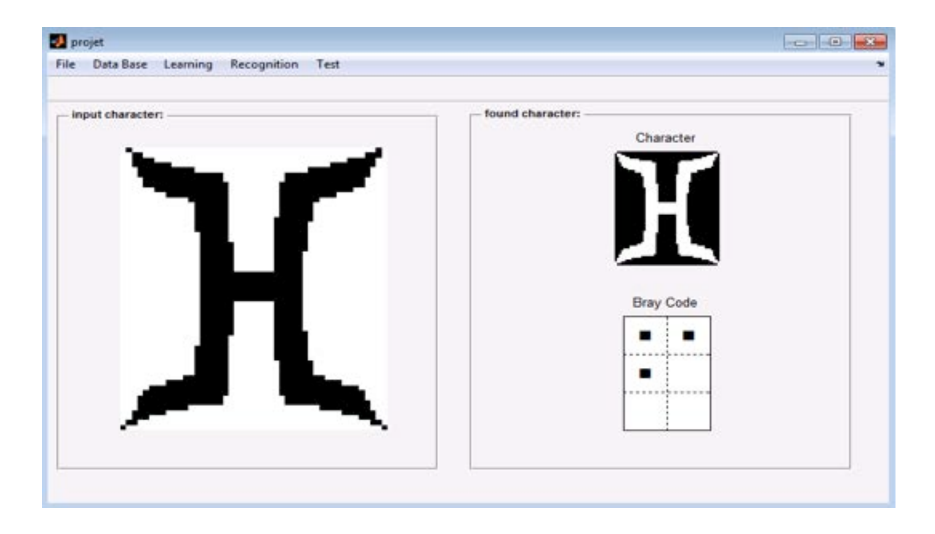


Figure 4 : Exemple 1 de résultat

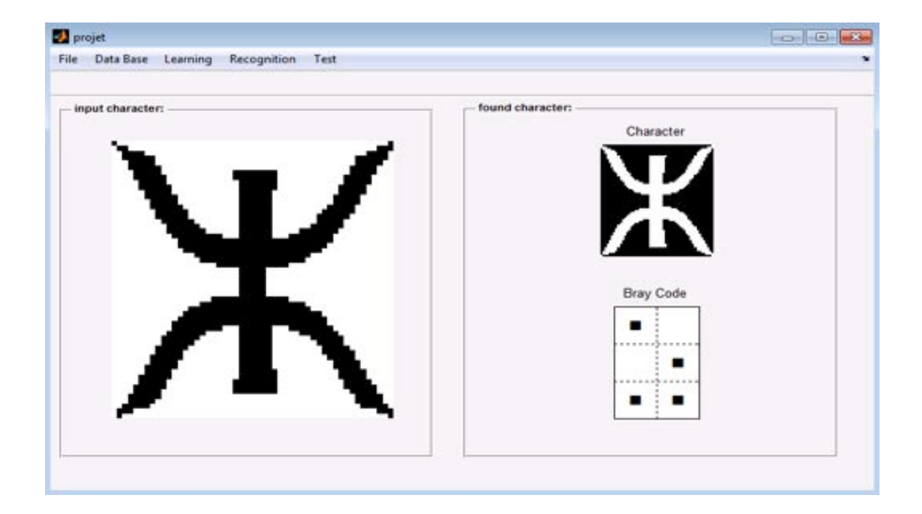


Figure 5 : Exemple 2 de résultat

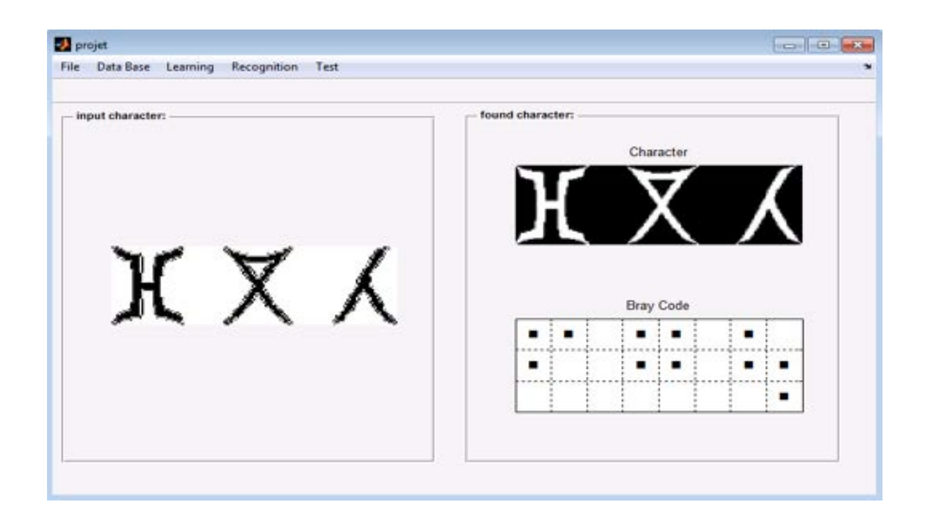


Figure 6 : Exemple 3 de résultat

La comparaison effectuée, entre les deux approches appliquées dans la phase d'extraction d'attributs, en se basant sur les réseaux de neurones comme méthode de classification, est illustrée par le Tableau 2.

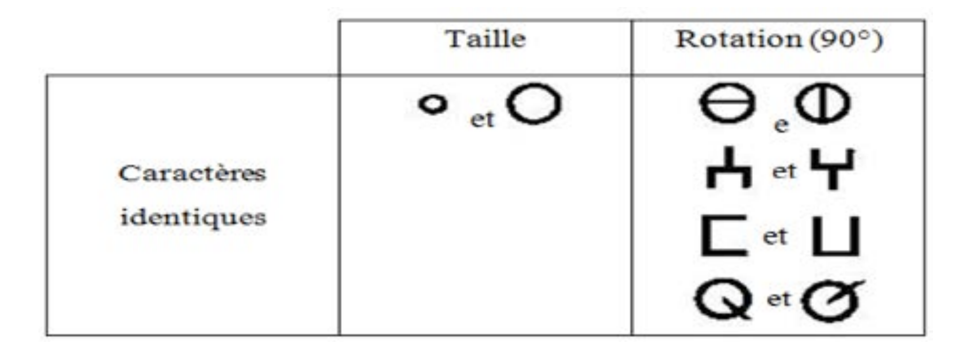
Classifications	Descripteurs	Taux de reconnaissance (%)	Taux d'erreur (%)	Temps d'exécution (s)
Réseaux de neurones	Moments de Zernike	80.61	19.39	859.92
	Moments de Legendre	81.21	18.79	853.16

Tableau 2 : Taux de reconnaissance, taux d"erreur et temps d'exécution

Les résultats du Tableau 2 montrent que :

- Le taux de reconnaissance calculé en utilisant les moments de Legendre est supérieur du taux de reconnaissance trouvé par les moments de Zernike.
- Le temps d'exécution, dans le cas d'utilisation des moments de Zernike, est supérieur du temps d'exécution des moments de Legendre.

Les descripteurs invariants posent des problèmes lors de la reconnaissance à cause de l'existence des caractères Tifinaghes qui se ressemblent (Tableau 3) en quelques situations (rotation et changement d'échelle).



	Taille	Rotation (90°)
Caractères identiques	et	e et et et

Tableau 3 : Caractères identiques

6. Conclusion

Le système de reconnaissance proposé a pour objectif d'afficher le code Braille correspondant à un texte écrit en Tifinaghe. Il comporte trois étapes : Prétraitement, l'extraction des attributs et la classification. Le taux de reconnaissance est environ de 80% en utilisant une base de données composée de 132 caractères dont 33 représentent les caractères de référence et 99 représentent ceux de test. Les erreurs sont dues à la rotation et au changement d'échelle. Tous les programmes sont implémentés sous le logiciel Matlab sur un PC duel core de RAM de 2 Giga Octet.

Référence

Alt F.L. (1962). Digital pattern recognition by moments. *Journal of the Association for Computing Machinery*, Vol. 9, pp. 240-258.

Amrouch M., Rachidi A., Elyassa M., Mammass D. (2010). Handwritten Amazigh character recognition based on hidden Markov models. ICGST-GVIP Journal, 10(5):11-18.

Chonga C.W., Raveendranb P., Mukundan R. (2004). Translation and scale invariants of Legendre moments. *Pattern Recognition*, Vol. 37, pp. 119-129.

El Ayachi R. (2012). Reconnaissance des formes par une classe d'invariants et applications à la reconnaissance des documents tifinaghes. Thèse de Doctorat, Faculté des Sciences et Techniques, Béni Mellal.

El Ayachi R., Bouikhalene B., Fakir M. (2012). Transformation de Fourier et moments invariants appliqués à la reconnaissance des caractères tifinaghe. *Revue électronique internationale en TI (eTI),* Numéro 6, ISSN 1114-8802.

El Ayachi R., Fakir M., Bouikhalene B. (2011). Recognition of tifinaghe characters using a multilayer neural network. *International Journal of Image Processing* (IJIP), 5(2):109-118.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011). Amazigh handwritten character recognition based on horizontal and vertical centerline of character. *International Journal of Advanced Science and Technology*, Vol. 33, pp.33-50.

Oujaoura M., El Ayachi R., Minaoui B., Fakir M., Bouikhalene B., Bencharef O. (2013). Invariant descriptors and classifiers combination for recognition of isolated printed tifinagh characters. *IJACSA Special Issue*, 3(2): 34-39, ISSghN 2158 – 107X, 2013.

Oujaoura M., Minaoui B., Fakir M., El Ayachi R., Bencharef O., (2014). Recognition of isolated printed tifinagh characters. *IJAC*, Vol. 85, n°1.

Teague M.R. (1980). Image analysis via the general theory of moments. *Journal of the Optical Society of America*, Vol. 70, pp. 920–930.

TICAM'14, p. 189-201

Statistical Methods for Amazigh OCR

Nabil Aharrane[1] **Karim El Moutaouakil**[2] **Khalid Satori**[1]
[1] Sidi Mohamed Ben Abdellah University
aharranenabil@gmail.com khalidsatori@gmail.com
[2] National School of Applied Sciences, Oujda
karimmoutaouakil@yahoo.fr

Abstract

The main purpose of this work is to develop an Optical Character Recognition system (OCR) of handwritten Amazigh characters employing a feature set of 79 elements based on statistical methods.

The feature set elaborated consists of 37 densities features and 42 shadow features based on a specific zoning to represent the Amazigh characters. In the recognition phase, we use the multilayer perceptron (MLP) as classifier.

In experiments evaluation, the accuracy observed on a large database of 24180 handwritten characters is 96, 13%. This evaluation not only verifies that the proposed approach provides a very satisfactory recognition rate but also shows a reasonable time during the test phase.

1. Introduction

In recent years, the recognition of characters handwritten remains one of the most popular problems due to its diverse applications such as address classification system, processing of bank check, indexing archives, documents analysis, etc. Therefore, much work has been achieved for many languages, an overview of the latest works in Optical Character Recognition (OCR) research can be found in (Peng *et al*., 2013).

Recently, researchers have begun to give attention to the Amazigh language OCR. In this context, various methods have been used based on: Hidden Markov Models (HMM) (Amrouch *et al*., 2012), Hough transformation (Oulamara & Duvernoy, 1988), neural approaches (El Ayachi *et al*., 2011; Es Saady *et al*., 2010; Gounane *et al*., 2011), geometrical and statistical methods (Bencharef *et al*., 2011; Es Saady *et al*., 2011; Fakir *et al*., 2011; Gounane *et al*., 2013), syntactical method rests on finite automata (Es Saady *et al*., 2011), moments features (Abaynarh *et al*., 2011; Oujaoura *et al*., 2013) and some hybrid methods (Amrouch *et al*., 2009; El Kessab *et al*., 2011; Moudni *et al*., 2013).

In this paper, we propose an OCR system based on a statistical approach with a new feature set. This latter creates, for each character, a set of features rests on decomposing the character

image under study in term of zones, and then we extract a vector of 79 components which are the densities features and the shadow features. After features extraction, due to its performance and its simple principle, we use the MLP with one hidden layer for recognition phase.

The rest of this paper is organized as follows: In Section 2, we present a description of the Amazigh language. The preprocessing description is given in Section 3, where we delineate all necessary steps to prepare the image to the next phases. A brief feature extraction state of art is presented in Section 4. Section 5 exposes the MLP architecture used for this work. Section 6 details our procedure to construct the feature set. In Section 7, we present some experimental results to evaluate our work. Finally, we conclude the paper with Section 8.

2. The Amazigh Language

The Amazighs are the indigenous people of North Africa, with their own language, culture and history. They are one of the most ancient peoples of humanity (CMA, 2006). Their presence in Tamazgha (North Africa) was more than 12000 years. The Amazigh language has existed since the earliest antiquity. It has an original writing system, Tifinagh, used and preserved to this day. In recent decades, all Amazigh groups have reclaimed this ancestral writing. Currently, the Amazigh language is spoken by about 30 million speakers in North Africa (from the oasis of Siwa in Egypt, to Morocco passing through Libya, Tunisia, Algeria, Niger, Mali, Burkina Faso and Mauritania).

In Morocco, where nearly 50% of people are Amazigh, the Amazigh language is divided into three regional varieties with Tarifite in North, Tamazight in Central Morocco and South-East and Tachelhite in South-West and the High Atlas (Ameur *et al.*, 2004).

Figure 1: Tifinagh characters adopted by IRCAM

The official introduction of the Amazigh language in the Moroccan educational system in 2003 involves the selection of a standard common language to teach. This task was accomplished by the "Royal Institute of the Amazigh Culture" (IRCAM) created in July 2001. Actually, the Tifinagh-IRCAM alphabet is based on 33 characters (Figure 1). In the Amazigh OCR field, one works only on 31 characters because ⴳⵯ and ⴽⵯ do not have a Unicode codes.

3. Preprocessing

In this section, we describe in details the preprocessing operations used to prepare the image for the next phases.

3.1. Binarization

The output of this operation is a binary image where black pixels represent the text and white pixels indicate the background. In this regard, several algorithms have been proposed in (Sezgin & Sankur, 2004).

In this work, we use the nonparametric and unsupervised Otsu's method (Otsu, 1979). This method performs an automatic thresholding that consists on maximizing the separability of the resultant classes in gray levels. It uses the zeroth and the first cumulative moments of the image histogram. The Otsu method gives good results and it still one of the most used thresholding methods.

3.2. Skew correction

The correction of the line skew consists in rectifying horizontally the oblique writing lines. Several methods are available in (Chin *et al*., 1997). The two most popular are the Hough transform and histograms projection. In this paper, we use the histograms projection, for its simplicity and its rapidity, based on scanning image according to directions D close to the horizontal, and counting the number of black pixels in these directions for each line. The quality of histogram is estimated by its entropy. The most probable direction is the one who maximizes this entropy. The document angle θ is that which corresponds to the histogram of maximum entropy. To correct this inclination, simply apply an image rotation with the angle θ.

3.3. Segmentation

The character segmentation is one of the most important steps in an Optical Character Recognition system (OCR). The objective is to decompose the image into a sequence of sub-images, each sub-image must contain a single character. For this, a line segmentation of the image is performed. Then each line is segmented into characters. A survey of methods and strategies in character segmentation is presented in (Casey & Lecolinet, 1996).

3.3.1. Line segmentation

In order to segment the image text into lines, we use the horizontal projection histogram. This method can distinguish between high density areas characterizing lines and low density areas indicating the space between lines.

3.3.2. From lines to characters

Since Amazigh writing, handwritten or printed, is never cursive, character extraction from each line becomes easy. In this context, we used vertical projection histogram. Characters correspond to areas of high density in the histogram.

3.4. Normalisation

The segmentation process produces isolated characters in different size. To solve this problem, we preceded to the normalization. This latter consists on resizing all characters to a common size. In this work, due to its zooming quality, we used a spline-based algorithm (Muñoz *et al.*, 2001) to resize all characters to a size of 30x30. This optimal spline-based algorithm for the enlargement or reduction of digital images can be realized through a new method of finite differences by calculating the scalar products with analysis functions that are B-splines of any degree. This algorithm achieves a reduction of artifacts such as aliasing and blocking and a significant improvement of the signal-to-noise ratio.

4. Feature extraction

The feature extraction step is a very important operation for a handwriting recognition system. It aims to select the most relevant information identifying each character to form a feature set.

In the literature, many feature extraction methods have been applied for OCR systems. (Arica *et al.*, 2001) categorize them according to their type as follows:

- Global Transformation and Series Expansion;
- Statistical features;
- Geometrical and topological features.

In this paper, after the preprocessing phase, we use a feature set based on statistical methods. This latter is based on the decomposition of the image characters that are segmented into several zones according to different directions. Then, their density and their length of projections are calculated. A detailed description of our approach is given in Section 6.

5. Character classification

After feature extraction of each segmented and normalized image, we use the resulting vector, consisting of 79 components, characterizing each character for learning and testing.

For this, several classification approaches are used in the field of handwriting recognition. According to (Jain & Jianchang, 2000), recognition techniques and text classification are grouped into four main categories:

- Pattern matching methods using correlation and distance measure;
- Statistical methods based on discriminant functions;
- Structural and syntactic methods employing rules and grammars;
- Neural networks classifiers.

In this paper, we use the MLP, one of the great families of neural networks, which use a supervised learning method called backpropagation (De Villiers & Barnard, 1992). The MLP used contains three layers:

- The input layer that consists of 79 nodes for the extracted vector ;

- The output layer with 31 nodes to distinguish the 31 classes which represent the number of the studied characters in the Amazigh language;
- One hidden layer in which the number of nodes is chosen experimentally.

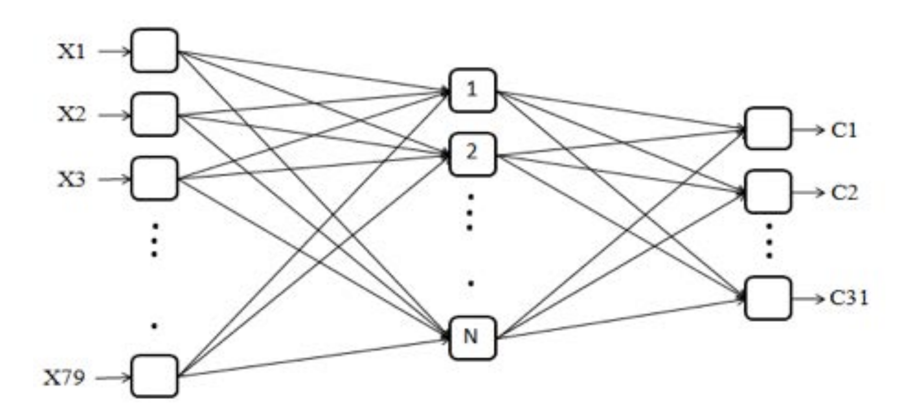


Figure 2: Architecture used of Multilayer perceptrron.

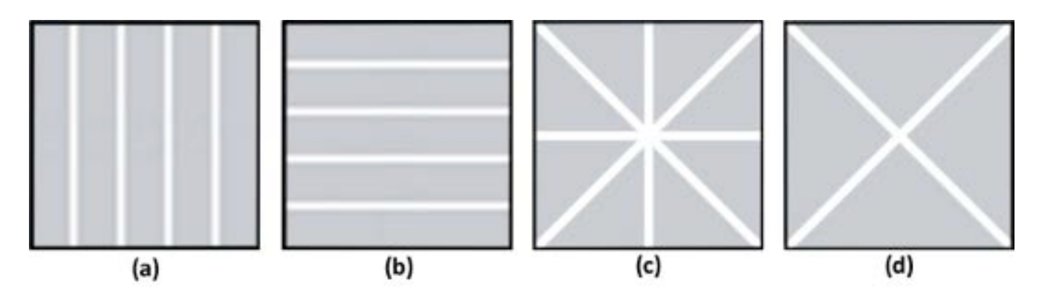
It should be noted that using one hidden layer is sufficient to solve a non linear complex problem, and that the choice of the number of hidden layer is still a challenging issue (Karsoliya, 2013).

6. Our approach

The choice of relevant features influences largely the performance of the character recognition system. In this context, we develop a feature set to provide a description that can characterize each character. This feature set consists of two subsets: The first one is generated by dividing the image into different overlapped zones. Then, we compute the densities of black pixels in each zone. Concerning the second subset, we calculate the shadow features (Basu *et al*., 2005) in different zones, where shadow features are the lengths of projections on different sides of the considered zones. The resulting feature set contains 79 components which 37 come from the first subset and the remainder is obtained by the second. To implement this method, the sizes of characters images under study are all resized to 30 × 30. In this section, we present in details the different stages to construct our feature set.

6.1. Density features

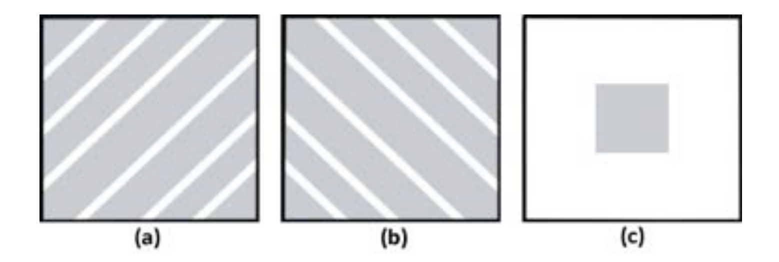
To create the first feature subset, we carry out different decompositions of the character image. Furthermore, the density of foreground pixels is calculated in each zone to obtain 37 features. We obtain the density of each zone by dividing the number of black pixels by the total number of pixels in this zone.



(a) Decomposition into 5 vertical equal zones;
(b) Decomposition into 5 horizontal equal zones;
(c) Decomposition into 8 octants;
(d) Decomposition into 4 quadrants.

Figure 3: First decomposition of character image

The first decomposition, as shown in the Figure 3, consists in dividing the character image vertically (Figure 3.a) and horizontally (Figure 3.b) to five equal zones, then to 8 octants (Figure 3.c) and in last to 4 quadrants (Figure 3.d).



(a) Left diagonal decomposition;
(b) Right diagonal decomposition;
(c) 10x10 middle zone.

Figure 4: Second decomposition of character image

The second decomposition is obtained by dividing the character image to 7 diagonal zones in both left and right directions (Figures 4-a, 4-b), then considering only the middle zone which size is 10x10. Diagonal features increase the recognition accuracy and reduce the misclassification. The middle zone was added to distinguish between some resembling Amazigh characters.

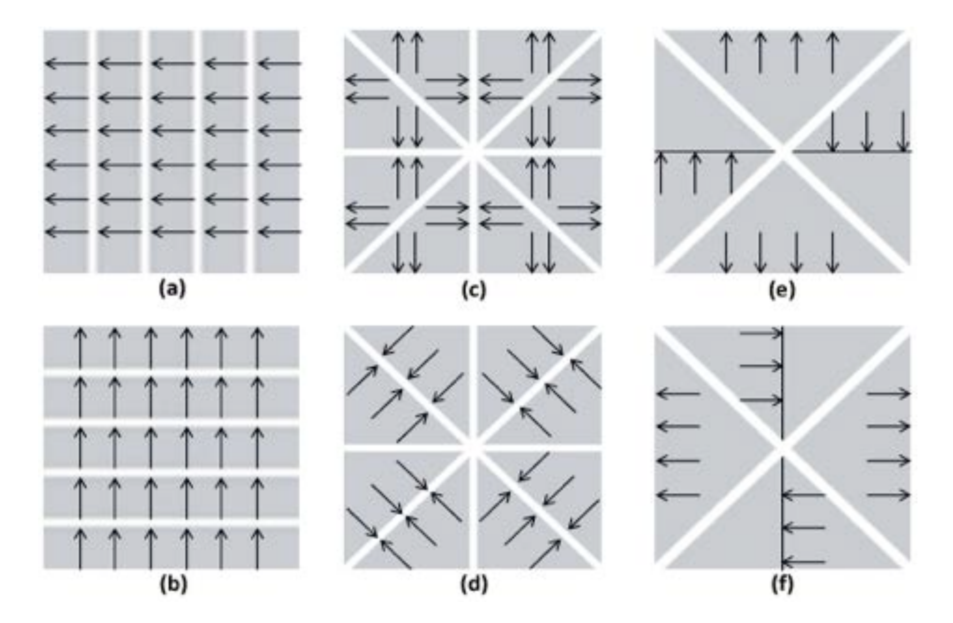
Following these decompositions, we obtain 37 zones and we calculated the density for each zone.

6.2. Shadow features

The 42 shadow features are obtained from the first decomposition of the rectangular boundary enclosing the character image (Figure 3).

We calculate the 10 lengths of horizontal and vertical projections (Figures 5-a, 5-b), 24 lengths of projections on each of the three sides of each octant (Figures 5-c, 5-d), and the 8 horizontal and vertical projections of each quadrant (Figures 5-e, 5-f).

Each calculated value must be normalized by dividing it on the maximum possible length of projections on the corresponding side.



(a) Shadow features of 5 vertical zones;
(b) Shadow features of 5 horizontal zones;
(c) Horizontal and vertical shadow features for each octant;
(d) Diagonal shadow features for each octant;
(e) Vertical shadow features for each quadrant;
(f) Horizontal shadow features for each quadrant.

Figure 5: Decomposition for shadow features

7. Experimental results and discussions

7.1. Database

In order to evaluate the performance of the proposed OCR system, the AMHCD database (Es Saady *et al*., 2011) was used as a source of training and test. The database consists of 25740 isolated Amazigh handwritten characters produced by 60 writers who wrote 13 samples of each 33 classes. As mentioned in section 2, researchers work only on 31 characters excluding characters ⴳⵯ and ⴽⵯ. So, experimentations were carried out on 24180 characters where 4030 were for training and the rest (20150) for test.

7.2. MLP

The fully connected three-layer perceptron neural network was trained using a sigmoidal activation function, a learning rate of 0.1, a momentum of 0.25 and all weights were randomly initialized in the interval [-0.7,0.7]. Several runs of backpropagation algorithm with 4030 epochs were performed for different architectures by varying the number of nodes in the hidden layer. The runs were executed in a compatible HP, Intel (R) Core (TM) Duo CPU 1.4 GHz, and 2 GB of RAM through Java. Table 1 shows the results obtained in tests with the chosen architectures using the feature set developed in this work.

Number of hidden nodes	Accuracy on test samples (%)
55	95,25
60	95,55
65	95,52
70	95,83
75	95,74
80	95,81
85	95,93
90	96,13
95	96,11

Table 1: Recognition rate for different Number of hidden nodes

It should be noted that the learning phase took nearly 2 hours and it is executed only once and weights are stored in a file, simply load the weights before performing recognition.

7.3. Discussion

According to the experiences, the best recognition performance of MLP is obtained when the number of hidden neurons is set to 90. In our experiences, we did not exceeded 95 neurons in the hidden layer to select a number of hidden neurons that provides a compromise between performance and the time taken in the recognition. Thereby, we opted for the 79-90-31 architecture for our MLP which allowed us to achieve a recognition rate of 96.13% and a reasonable time in the recognition phase which is 4 milliseconds for each character.

Basing on this architecture, we computed the individual accuracy for each class of the Amazigh characters in the test data, the results obtained are shown in Table 2.

Character	Accuracy (%)
ⴰ	100
ⵣ	84.0
ⴻ	97.7
ⵔ	94.15
ⵜ	97.54
ⵢ	95.23
ⵓ	97.23
ⵄ	96.92
ⵃ	98.46
ⵛ	98.77
ⵉ	96.62
ⴱ	93.85
ⴷ	99.38
ⴼ	97.85
ⴳ	95.38
ⵀ	97.85

Character	Accuracy (%)
ⵊ	98.15
ⴽ	97.38
ⵍ	97.38
ⵎ	98.92
ⵡ	99.85
ⵅ	92.46
ⴳⵯ	98.15
ⵖ	97.69
ⵙ	98.15
ⵏ	99.85
ⵥ	86.0
ⵚ	96.15
ⵟ	90.77
ⴹ	94.62
ⵇ	93.69

Table 2: Individual recognition rate for each character

The obtained results show that some characters have a relatively low recognition rate compared to others, especially for characters yaz (ⵣ) yazz (ⵥ) and yatt (ⵟ). The misclassifications are due to 2 factors. The first one is the structural similarity of some characters. Table 3 shows the main confusions between characters.

Character	Classified as	Number of confusions (%)
ⵥ	ⵣ	9.38
ⵣ	ⵥ	7.23
ⵟ	ⴹ	5.38
ⵣ	ⵅ	3.38
ⵙ	ⵔ	3.38
ⵅ	ⵣ	3.07
ⴹ	ⵎ	2.46

Table 3: Main confusions between characters

The second factor is the bad writing of some characters in the database whose classification is difficult even for a human operator. Figure 6 illustrates some badly written letters in the database.

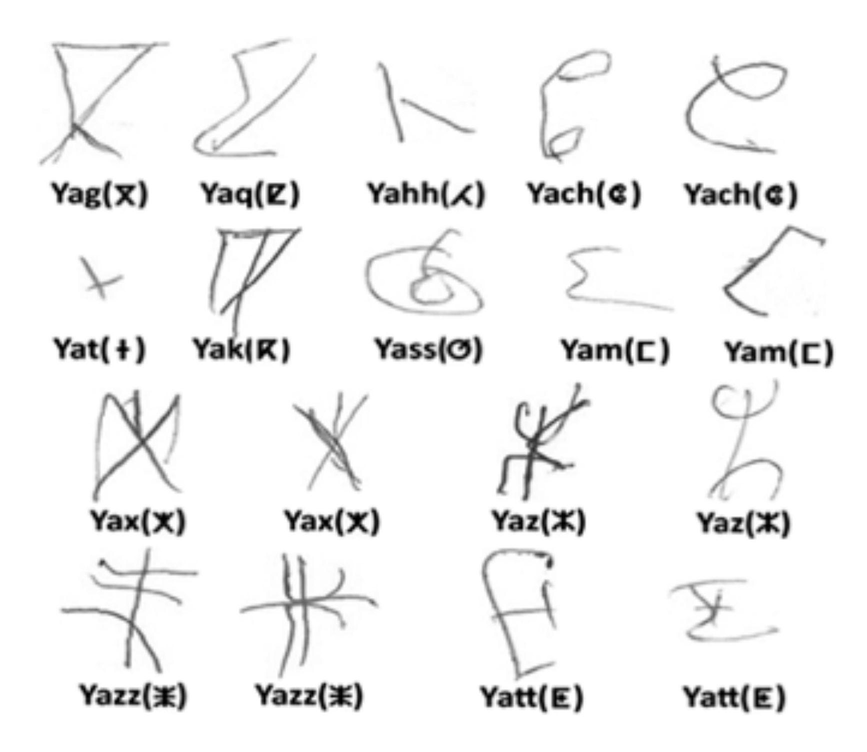


Figure 6: Some characters badly written in the database

To showcase our method, we compared our obtained results with those of (Amrouch *et al*., 2012), based on the same database. It should be noted that the latter work reports the highest recognition accuracy of about 97.89% in handwritten Amazigh language OCR by using continuous HMMs. Amrouch *et al*. used 66.67% of the database in the training phase that can cause an overfitting. Furthermore, they tested their system on, only, 33.33% of the database. As for us, we have used 16.67% of the database to train the MLP and 83.33% to test it. This shows that our system is more intelligent than their system.

8. Conclusion

In this work, we have proposed an optical character recognition system of handwritten Amazigh characters employing a statistical approach to develop a new feature set. This one consists on calculating the densities and shadow features of each character by decomposing the image under study in term of zones. Basing on this latter, we have extracted a vector of 79 components to represent each character. Due to its performance, we have used the MLP in the recognition phase. Some experimental results are introduced.

According to the experimental analysis, we conclude that the chosen statistical features are useful features to describe Amazigh characters and recognition rate can be very satisfactory.

References

Abaynarh M., Elfadili H., Zenkouar L. (2011). Recognition of tifinaghe handwritten characters using moments for feature extraction. *Actes du 4ème atelier international sur l'amazighe et les TICs, sous le thème 'Les ressources langagières : construction et exploitation'*, pp. 345-356.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E., Souifi H. (2004). Initiation à la langue amazighe. Manuels n°1, pp. 9, Rabat : IRCAM.

Amrouch M., Es-Saady Y., Rachidi A., El Yassa M., Mammass D. (2009). Printed amazigh character recognition by a hybrid approach based on hidden Markov models and the Hough transform. *Multimedia Computing and Systems*, pp. 356-360.

Amrouch M., Es-saady Y., Rachidi A., El Yassa M. and Mammass D. (2012). Handwritten Amazigh Character Recognition System Based on Continuous HMMs and Directional Features. *International Journal of Modern Engineering Research*, 2(2):436-441.

Arica N., Yarman-Vural F.T. (2001). An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and reviews*, 31(2):216-232.

Basu S., Das N., Sarkar R., Kundu M., Nasipuri M., Basu D.K. (2005). Handwritten Bangla alphabet recognition using MLP based classifier. *Proceedings of the 2nd National Conference on Computer Processing of Bangla*, pp. 285-291.

Bencharef O., Fakir M., Minaoui B. (2011). Tifinagh character recognition using geodesic distances, decision trees & neural networks. *International Journal of Advanced Computer Science and Applications. Special Issue on Artificial Intelligence*, pp. 51-55.

Casey G., Lecolinet E. (1996). A Survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):690-706.

Chin W., Harvey A., Jennings A. (1997). Skew detection in handwritten scripts. *Proceedings of IEEE TENCON'97, IEEE Region 10 Annual Conference, Speech and Image Technologies for Computing and Telecommunications*, pp. 319-322.

Congrès Mondial Amazigh CMA (2006). Les Amazighs du Maroc : un peuple marginalisé. Comité des Nations Unies pour les Droits Economiques, Sociaux et Culturels Genève, pp 2.

De Villiers J., Barnard E. (1992). Backpropagation neural nets with one and two hidden layers. *IEEE Transactions on Neural Networks*, 4(1):136-141.

El Ayachi R., Fakir M., Bouikhalene B. (2011). Recognition of tifinaghe characters using a multilayer neural network. *International Journal Of Image Processing*, 5(2):107-118.

El Kessab B., Daoui C., Moro K., Bouikhalene B., Fakir M., (2011). Recognition of handwritten tifinagh characters using a multilayer neural networks and hidden Markov model. *Global Journal of Computer Science and Technology*, 11(15):12-20.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011). Amazigh handwritten character recognition based on horizontal and vertical centerline of character. *International Journal of Advanced Science and Technology*, Vol. 33, pp. 33-50.

Es Saady Y., Rachidi A., El Yassa M., Mammass D. (2011). AMHCD: A Database for Amazigh handwritten character recognition research. *International Journal of Computer Applications*, 27(4): 44-48.

Es Saady Y., Rachidi A., El Yassa M. and Mammass D. (2010). Printed Amazigh character recognition by a syntactic approach using finite automata. *ICGST-GVIP Journal*, Vol. 10, No. 2, pp. 1-8.

Fakir M., Bencharef O., Bouikhalene B. and Minaoui B. (2011). Tifinagh Character Recognition Using Riemannian Metric, SVM & Neural Networks. *International Journal of Advances in Science and Technology*, 2(6):1-9.

Gounane S., Fakir M., Bouikhalene B. (2011). Recognition of tifinagh characters using self organizing map and fuzzy K-Nearest neighbor. *Global Journal of Computer Science and Technology*, 11(15):28-34.

Gounane S., Fakir M., Bouikhalene B. (2013). Handwritten tifinagh text recognition using fuzzy K-NN and Bi-gram language model. *International Journal of Advanced Computer Science and Application, Special Issue on Selected Papers from Third International Symposium on Automatic Amazigh Processing (SITACAM'13),* pp. 29-32.

Jain A. K., Jianchang M. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*. 22(1), pp. 4-37.

Karsoliya S. (2012). Approximating number of hidden layer neurons in multiple hidden layer bpnn architecture. *International Journal of Engineering Trends and Technology*, 31(6):714-717.

Moudni H., Er-rouidi M., Oujaoura M., Bencharef O. (2013). Recognition of Amazigh characters using SURF & GIST descriptors. *International Journal of Advanced Computer Science and Application, Special Issue on Selected Papers from Third International Symposium on Automatic Amazigh Processing (SITACAM'13),* pp. 41-44.

Muñoz Barrutia A., Blu T., Unser M. (2001). Least-Squares image resizing using finite differences, *IEEE Transactions on Image Processing*, 10(9):1365-1378.

Otsu N. (1979). A Threshold selection method from Gray-Level histograms. *IEEE Transactions on Systems*, Man and Cybernetics, pp. 62-66.

Oujaoura M., El Ayachi R., Minaoui B., Fakir M., Bouikhalene B. (2013). Invariant descriptors and classifiers combination for recognition of isolated printed tifinagh characters. *International Journal of Advanced Computer Science and Application, Special Issue on Selected Papers from Third International Symposium on Automatic Amazigh Processing (SITACAM'13),* pp. 22-28.

Oulamara A., Duvernoy J. (1988). An application of the Hough transform to automatic recognition of Berber characters. *Signal Processing*, Vol. 14, pp. 79-90.

Peng X., Cao H., Setlur S., Govindarju V., Natarajan P. (2013). Multilingual OCR research and applications: An overview. *Proceedings of the 4th International Workshop on Multilingual OCR*, ACM, New York, NY, USA. Article No.1.

Sezgin M., Sankur B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146-168.

New Approach of Tifinagh Character Recognition using Graph Matching

Youssef Ouadid Mohammed Fakir
Brahim Minaoui Otmane Abouelala
Team Information Processing and Telecommunications, Department of Computer Science, Faculty of Science and Technology, Beni Mellal, USMS, Morocco

Abstract

Most existing methods of character recognition require lot of preprocessing. Also using methods of artificial intelligent (neuralnetwork, SVM ...) request many resources and time (time necessary for the learning phase and recognition). In this paper, we present a structural approach to recognize Tifinagh characters. In this approach, we use graphs representation as a model for offline Tifinagh character recognition. It allows extracting the basic form of an independent letter and recognizes it. The proposed approach shows its effectiveness against problems such as size and orientation resulting performance gain, also in term of its speed.

1. Introduction

Optical Character Recognition (OCR) is a computer procedure that converts an image of printed text into a text file. This means getting the text in image of printed text and save it in a readable text file for treatment and enrichment. The images can come as appliances of online or offline devices. Online devices work with a pen and include displays, tablets and graphics tablets. These tablets can provide temporal order of the points that represent the lines of text. Some tablets provide additional information, including speed and pressure. Whereas, offline devices include scanners flat type, manual and paper. They make an image substantially in the form of raster pixels.

Amazigh is a Chamito-Semitic language derived from ancient Berber. It includes a variety of dialects present from Morocco to Egypt, passing through Algeria, Tunisia, Mali, Niger and Libya. Several research studies have been performed on the computerization of Tifinagh characters. This research is concentrated in Morocco and Algeria. In Morocco there are three research centers: Sultan Moulay Slimane University in Beni Mellal, Ibn Zuhr University in Agadir, and Royal Institute of Amazigh Culture (IRCAM) in Rabat.

Lately, a lot of research has been done on Tifinagh OCR. In their Articles, Essaady *et al.* (2010) adopted an analytical approach in which they used finite automata as a method of recognition and Freeman coding as feature extraction method. Bencharef *et al.* (2011) used a Riemannian metric as a feature. These metrics descriptors are reliable toward change scale, noise and geometric distortions. El Ayachi *et al.* (2010; 2011) proposed two methods of

recognition. Both methods use the multilayer neuron networks and dynamic programming for the classification, and Moments invariants Walsh transform for feature extraction. In their work, neural networks combined with Walsh transform showed fairly interesting results. Oujaoura *et al*. (2014) have merged several feature extraction methods (Hu Moments, Zernike Moments, Legendre Moments, Texture, GIST, Walsh) and several classification methods (Nearest Neighbor, Multiclass SVM, Bayesian Networks, Neural Networks), as a result they got an excellent recognition rate but a large processing time. All the methodologies adopted in this work give good results, but it require lot of preprocessing which makes these systems long enough to deal with a set of scripts.

Hence, the idea of Hebb (Arrivault, 2002-a) that any form is apprehended by the human visual perception, not as a whole, but in parts. Furthermore, the organization and the relative spatial positions of its parts play an important role in learning and recognition. Structural descriptions are based on a decomposition of forms into simple elements. In our system of Tifinagh OCR, we adopt a structural approach in which graphs are used as a model representation.

2. Recognition system Tifinagh characters

2.1. Characteristics of Tifinagh

Tifinagh alphabet is used by Berbers, mainly Tuaregs. It was formerly a consonantal alphabet. This alphabet has undergone inevitable modifications and variations since its origin until today. Tifinagh characters are normally written from left to right and vertically from top to bottom. Figure 1 illustrates Tifinagh characters adopted by IRCAM. It is composed of thirty-three characters, representing consonants and vowels.

Figure 1: officials Characters Tifinagh-IRCAM

2.2. Character recognition system

A Character Recognition System (CRS) accepts the output from an online or offline device as a data input, then, processes it and produces an understandable data output. The system is based on three major steps: preprocessing, feature extraction and classification, as shown in Figure 2.

Preprocessing phase aims to reduce the amount of data to keep only relevant information. It is called low-level treatment. The extraction of the primitive phase allows making from the image representation a synthetic description of the pattern to be recognized in a multidimensional space.

The classification phase is the phase of the character or word recognition. This step implies the existence of a learning database. The description of the character to be recognized is compared with the descriptions of the base character.

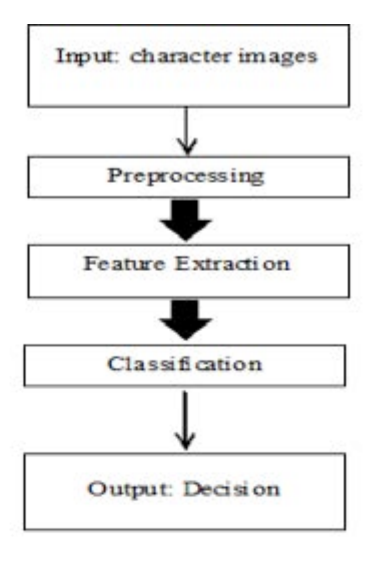


Figure 2: Character recognition system

2.3. Adopted system for Tifinagh character recognition

Figure 3 shows the adopted system for recognizing Tifinagh characters, and represents a set of methods and algorithms we use for each step.

2.3.1. Pre-processing

In order to automate the handwriting recognition, we must prepare images to be processed. Preliminary steps, which are necessary for recognition, are binarization, normalization and skeletonization.

a. Binarization

Binarization is the first step of the preprocessing phase. It consists in converting the scanned image into a binary image. However, binarization is an operation that produces two classes of pixels. In general, they are represented by black pixels and white pixels.

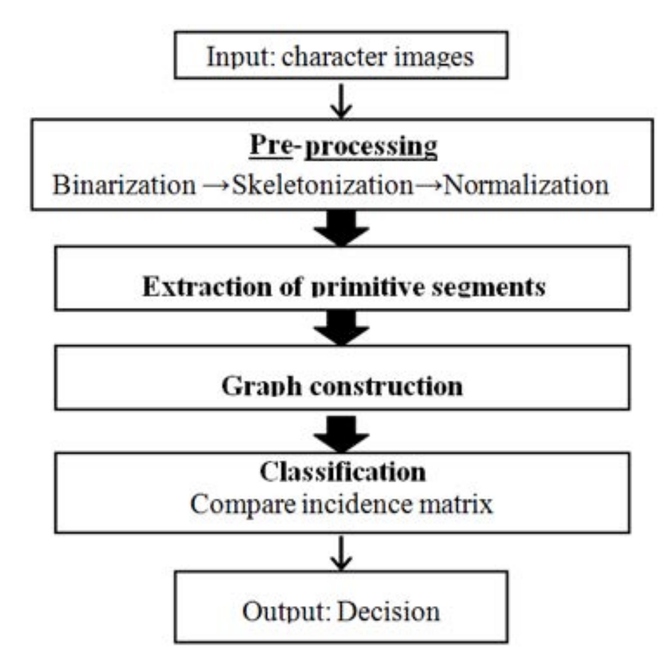


Figure 3: The adopted system for character recognition

b. Normalization of the position:

The goal is to improve the processing time by removing unwanted areas. To do that, we calculate the horizontal and vertical histograms to detect the first white pixel at the top, the bottom, the left and the right.

c. Skeletonization (Figure 4)

It is used to obtain a thickness equal to one of the write line and thus be reduced to a linear writing. The skeleton must preserve the shape, connectivity, topology and end of the route. Furthermore, it should not introduce parasitic elements.

We have adopted the algorithm of Zhang and Wang (1994) due to its strength and speed. This is a parallel algorithm in a single iteration that produces perfectly skeletons 8-connected and which operates the collisions.

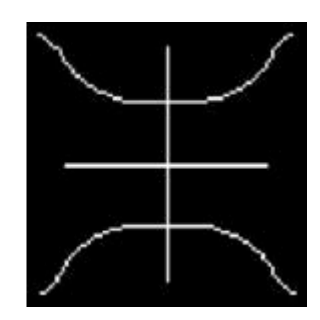


Image binarized *Image skeletonized*

Figure 4: Example of skeletonization

This algorithm removes all pixels that are not part of the skeleton. For pixel p1, we use 4*4 neighbors (Figure 5).

P10	P11	P12	P13
P9	P2	P3	P14
P8	P1	P4	P15
P7	P6	p5	P16

Figure 5: Pixel P1 and its neighbors

A pixel p1 removed if all of the following conditions are satisfied:

- P1=1.
- 2≤B(p1)≤6 where B(p1) is number of P1 white neighbors.
- A(P1)=1 where A(p1) is number of 01 pattern starting from P2.
- P2*P4*P8=0 or P11=1.
- P2*P4*P6=0 or P15=1.

2.3.2. Extraction of primitive segments

Once the skeleton of a character is obtained, we need to extract the primitive segments. These segments are geometrically as simple as possible. They constitute a grapheme which is schematically represented by a skeleton. The extracted singular points (Figure 4) can be divided into three types of minutiae points (Arrivault, 2002-b):

- End Points.
- Intersection Points.
- Inflection Points.

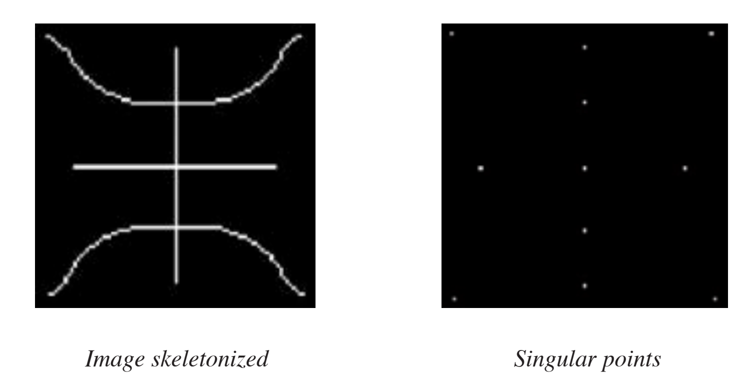


Image skeletonized *Singular points*

Figure 6: Example of singular points

a. Intersection and end points detection

This step does not pose a major problem since it's based on the study of the neighbors of each point. End points have only one 8-neighbor and the intersection points at least three. One method is to scan the image and check if the neighbors of each point correspond to a boolean mask characteristic of an intersection or end. Thus, the end points correspond to the following masks:

$$\text{With rotation of } 45^\circ \begin{cases} I^1 = \begin{matrix} 1 & 0 & 1 \\ x & 1 & 0 \\ 0 & x & 1 \end{matrix} \\ I^2 = \begin{matrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ x & 1 & x \end{matrix} \\ I^3 = \begin{matrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{matrix} \end{cases}$$

Where the symbol x represents a 1 or a 0, and the intersection points are obtained, meanwhile, with the following masks:

$$\text{With rotation of } 45^\circ \begin{cases} F^1 = \begin{matrix} 1 & 0 & 0 \\ x & 1 & 0 \\ 0 & 0 & 0 \end{matrix} \\ F^2 = \begin{matrix} 1 & x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix} \end{cases}$$

b. Inflection points Detection

After detecting the end points and intersection, the skeleton may be divided into a set of primitives whose concavity is not necessarily strict. The inflection points are calculated so primitive by primitive. By definition, an inflection point is where the curvature of the curve vanishes. A simple way is to use an interpolation then find places where D=(x" * y' - x' * y")=0. The De Casteljau algorithm (Arrivault, 2002-b) is quite appropriate as smooth breaks. However, the use of all points of the control points (points where the curvature of the curve vanish) as primitive, leads to interference detections. Using a polygonization Ramer (Arrivault, 2002-b) leads to a smoothing of the curve. With this, the number of control points is limited, thereby improving the results.

2.4. Graph construction

In this step, our goal is to give a structural description to our character using graphs. Graphs are considered as a set of very powerful structural data for the representation of objects.

A graph G is a pair G = (N, A), which represents N nodes where the vertices and A represents the arcs or edges. In a graph representation, vertices represent objects or parts of objects. The arcs represent the relationships between the objects or parts of objects. They can be oriented or non-oriented (that is to say a vertex directed to another). The graph associated with the matrices, such as matrices of impacts, contains information about the connectivity graph.

An impact matrix M is a matrix of size n * m, such that:

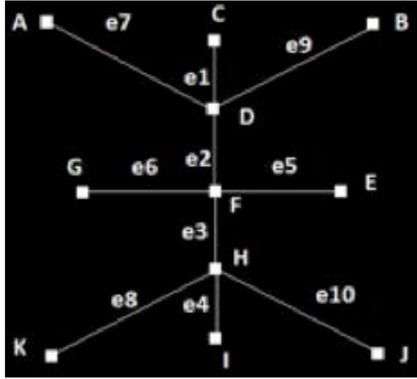
$$\begin{cases} n=|N| \text{the number of nodes} \\ m=|A| \text{the number of edges)} \end{cases}$$

An element $e_{i,j}$ (i= {1… n} and j= {1… m}) of the matrix M can has two values:

$$\begin{cases} e_{i,j} = 1 \text{ if the arcj is incident to top i} \\ e_{i,j} = 0 \text{ if no} \end{cases}$$

The degree of the graph is the number of edges incident to vertices, this means that for an undirected graph the degree is twice the number of arcs as each arc is incident to two vertices. The degree of a vertex is the number of arcs incident to the top.

Convert the image into a graph is critical because the graph should represent the characteristics of the letter to a large extent. The distance between two nodes must be high enough to represent the characteristics of the letter properly and also small enough to not take a lot of memory. This task has been carried out in extracting singular points. From these points, we can construct the appropriate graph (Figure 5).



Original image *Image graph representation*

Figure 7: Example of Image graph representation

3. Classification and Results

The training set consists of the total images incidence matrix reference (ShobhaRani, 2013) (see Figure 7 and Figure 8). The similarity test is made by comparing the incidence matrix of the test image with those of the reference images. A first step is to verify the number of nodes which means that we are using an exact graph matching method. This method affects the recognition rate but the CPU time is good.

	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10
A	0	0	0	0	0	0	1	0	0	0
B	0	0	0	0	0	0	0	0	1	0
C	1	0	0	0	0	0	0	0	0	0
D	1	1	0	0	0	0	1	0	1	0
E	0	0	0	0	1	0	0	0	0	0
F	0	1	1	0	1	1	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0
H	0	0	1	1	0	0	0	1	0	1
I	0	0	0	1	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	1
K	0	0	0	0	0	0	0	1	0	0

Figure 8: Example of incidence matrix

In this experience, 2200 character images from the Tifinagh-IRCAM base were used to form the training set, and 1100 character images were used to test the performance of identification. Some results on the number of states are listed in the table below.

These results demonstrate interesting in terms of CPU time compared to the works carried on this kind of characters. This comes from the fact that we only use singular points to represent character by graph. Also, it shows effectiveness against problems such size and orientation by using incidence matrix as a way to identify characters. Recognition rate can be improved by adding other features such as height and width of characters, also by improving the singular point detection.

Adopted method	Training data size	1100	2200
	Test database size	550	1100
	Recognition rate (%)	89.21	91.01
	Error rate (%)	10.79	8.99
	Training time (s)	300.25	400

Table 1: Recognition rate, error rate and training time

4. Conclusion

In this paper, we used graphs as representation model for recognition of printed Tifinagh characters. We used singular points as information to build character than extract incidence matrix and use it as the main feature to recognize character.

We have presented experimental results performed on the database of printed Amazigh characters. The results obtained are 91.01% with a computation time from 400 seconds using a database of 2200 characters.

The work done is a first step for several perspectives. We try to improve our recognition rate by changing the graph model for a better representation keeping the speed of the system. We try to extend our approach to the recognition of printed handwriting characters, the recognition of words, sentences and texts.

References

Arrivault D. (2002). Apport des graphes dans la reconnaissance non-contrainte de caractères manuscrits anciens. PhD Thesis, University of Poitiers, France, pp. 61.

Arrivault D. (2002). Apport des graphes dans la reconnaissance non-contrainte de caractères manuscrits anciens, PhD Thesis, University of Poitiers, France, pp. 74-77.

Bencharef O., Fakir M., Idrissi N., Bouikhalen B., Minaoui B. (2011). Application de la géométrie riemannienne à la reconnaissance des caractères Tifinagh. *SITACAM* Agadir-Morocco, May 06-07 2011, pp. 179-188.

El Ayachi R., Fakir M., Bouikhalene B. (2011). Recognition of tifinagh characters using dynamic programming & neural network. Chapter in Document Recognition and Understanding, ISBN 978-953-307995-0, INTECH.

El Ayachi R., Fakir M., Bouikhalene B., Safi S. (2010). Offline printed Amazigh scripts recognition. *JATIT*, 20(2).

EsSaady Y., Rachidi A., El Yassa M., Mammass D. (2010). Printed Amazigh character recognition by a syntactic approach using finite automata. *ICGST-GVIP Journal*, 10(2):1-8.

Oujaoura M., Minaoui B., Fakir M., El Ayachi R., Bencharef O. (2014). Recognition of isolated printed tifinagh characters. *International Journal of Computer Applications*, 85(1):1-13.

ShobhaRani N. (2013). Optical character recognition in application with graph theory and matrix processing. *IJCSRT*, pp. 16-18.

Zhang Y.Y., Wang P.S.P. (1994). A modified parallel thinning algorithm. *International Journal of Pattern Recognition and Artificial Intelligence*, 1994.

TICAM'14, p. 213-217

Logiciel de reconnaissance optique des Caractères du Tifinaghe : Démonstration

Youssef Ait Ouguengay **Younes El Affani**
Institut Royal de la Culture Amazighe, Rabat, Maroc
{aitouguengay, elaffani}@ircam.ma

Abstract

Optical character recognition (OCR) means computer process of the translation of printed text images typescripts to text files. Until a few years before, this area was not more advanced. However, the ROC has evolved with the advent of free solutions such Tesseract and Ocropus. The demonstration that we present in this paper, shows results of the integration of Tifinagh, the Amazigh alphabet, into the Tesseract recognition engine. Obtained results show an acceptable accuracy of 82.77%, for a first OCR softawre on Tifinagh, measured as the percentage of correct characters.

Résumé

La reconnaissance optique des caractères (ROC) désigne les procédés informatiques pour la traduction d'images de textes imprimés ou dactylographiés en fichiers de texte. Jusqu'à quelques années avant, ce domaine n'était pas des plus avancés. Cependant, la ROC a évolué et a été démocratisé avec l'apparition des solutions libres telles *Tesseract* et *Ocropus*. La démonstration que nous présentons, montre le résultat de l'intégration du Tifinaghe, alphabet amazighe, dans le moteur de reconnaissance des caractères Tesseract. Les résultats obtenus donnent une fiabilité moyenne de 82,77% qui est le pourcentage des caractères reconnu correctement.

1. Introduction

La reconnaissance optique de caractères (OCR) est une tâche de classification complexe dans lequel des images de texte sont analysées, en traduisant le texte dans les images en un format éditable. Cela peut être utile dans un large éventail de domaines, depuis la lecture du texte à partir des documents scannés vers le tri automatique du courrier. L'OCR peut être mis en œuvre sur un ensemble de données – nettes (dactylographié, imprimé) jusqu'aux textes bruités (manuscrits). Ces données constituent l'entrée qui peut rendre le travail de l'OCR relativement facile ou extrêmement difficile et sujette aux erreurs.

L'objective de travail actuel est de présenter un logiciel de reconnaissance optique des caractères amazighes ayant un degré de fiabilité acceptable. Le présent article fait, en fait, suite à un travail antérieur (Ait Ouguengay, 2011) où la construction du système comprenant les étapes nécessaires d'entrainement du moteur de reconnaissance ont été détaillées.

Dans la section suivante, nous attelons à l'architecture générale du logiciel, ensuite nous introduisons l'interface graphique du logiciel avant de présenter une synthèse des résultats de démonstration de la reconnaissance des caractères tifinaghes affichés par le logiciel ainsi conçu.

2. Intégration de l'interface graphique au moteur de reconnaissance Tesseract:

Le moteur de reconnaissance optique *Tesseract* est servi en mode ligne (Smith, 2007) et ne dispose pas d'une interface graphique intégrée. Disponible sur plusieurs distributions Linux, nombreuses adaptations de l'interface d'utilisateur graphique ont été proposées pour le moteur Tesseract. Cependant,' rares sont les propositions adoptées pour le système d'exploitation Windows, dont nous avons retenu l'adaptation multiplateforme VietOCR[1].

Ainsi, pour avoir le moteur opérationnel pour la numérisation d'un script donnée, il fallait installer plusieurs parties: le moteur lui-même avec la version adéquate, l'interface graphique et les données d'entrainements pour le langage cible. Au terme de notre travail, nous avons procédé à la compilation de toutes ses composantes en un seul extracteur de fichier (.exe) en utilisant un système d'installation programmable[2].

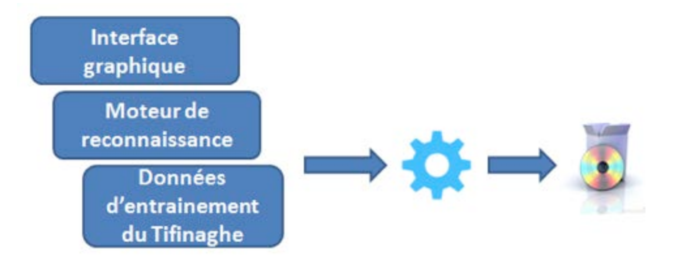


Figure 1 : Compilation des composants de logiciel OCR

L'installeur tel qu'il est compilé prend en charge de manière automatique l'installation des différentes parties du système OCR.

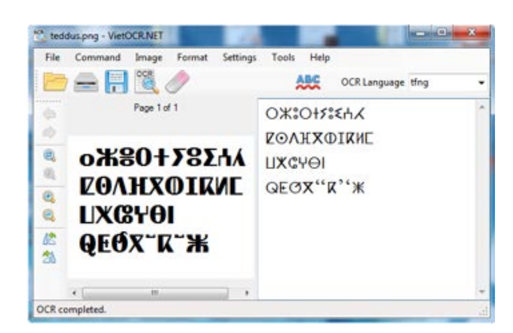


Figure 2 : Aperçu de l'interface graphique (Veit OCR .NET)

1 http://vietocr.sourceforge.net/
2 Inno setup, http://jrsoftware.org/isinfo.php

3. Synthèse des résultats de l'OCR :

Le logiciel OCR a été exécuté à travers une série de tests. Au total 18 fontes ont été mises en œuvre dans la préparation des données d'apprentissage et de tests. Initialement, le système a été entrainé par 14 différentes fontes, et de façon standard 70% de données ont été utilisées pour l'entrainement du moteur de reconnaissance optique et 30% pour le test. Les résultats de la reconnaissance des caractères pour les données d'entrainement sont décrits dans la figure 3.

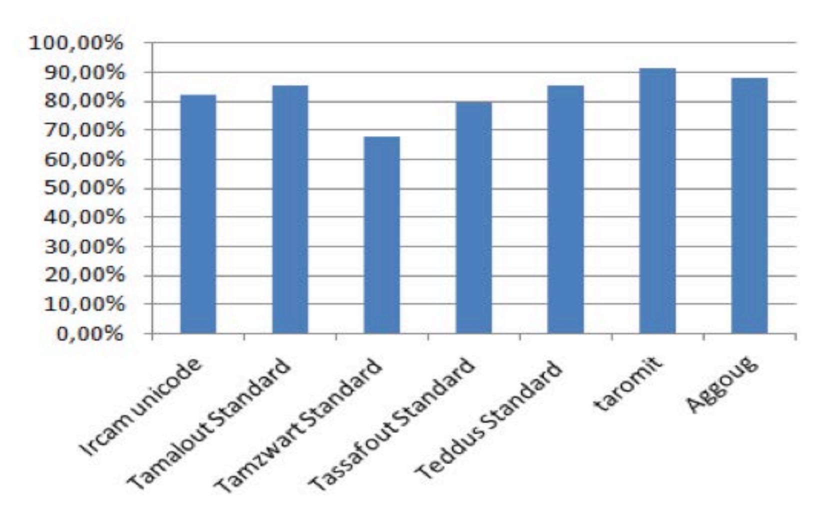


Figure 3 : Taux de reconnaissance pour les fontes d'apprentissage

Il est à noter que le système construit enregistre un taux de fiabilité important. Le taux de reconnaissance moyen dépasse 80 % aussi bien pour les données d'entrainement que pour ceux de test comme il est montré dans la figure 4 ce qui démontre une bonne tendance d'indépendance par rapport aux fontes, en tenant en compte que les fontes utilisées dans le corpus du test sont d'une complexité moyenne de même niveau que les fontes d'entraînement.

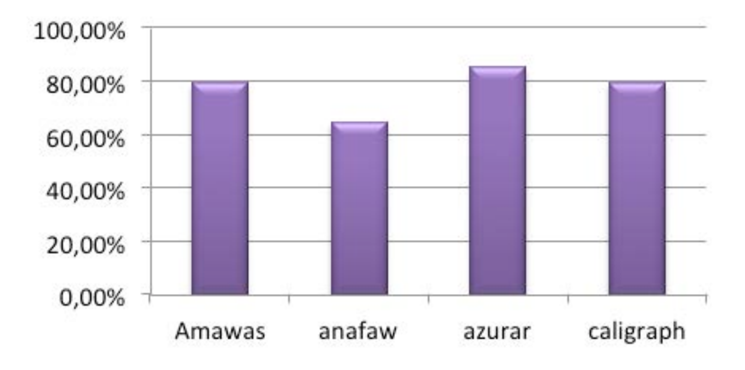


Figure 4 : Taux de reconnaissance des fontes du test

Pour chaque fonte, la mesure d'évaluation du produit peut être exprimée par le taux d'erreur de caractère (Rice, 1995 ; Minesh, 2016) (Character Error Rate-CER) qui est définis comme suit :

$$CER = \frac{\text{total d' erreurs de reconnaissance des caractéres}}{\text{total de caractéres}} \text{ x } 100$$

Le total d'erreurs de reconnaissance de caractère est mesuré d'une façon générale par la distance de Levenshtein[3] entre l'entrée de l'OCR et la sortie texte. Les erreurs de reconnaissance de caractères peuvent être :

- une confusion, en remplaçant un caractère par un autre, si les caractères sont morphologiquement proches (par ex. « ⵔ, 0 ») ;
- une suppression, en ignorant un caractère, considéré comme un bruit de l'image,
- un ajout, en dédoublant un caractère par deux autres dont la morphologie de leurs formes accolées peut être proche du caractère (par ex. «ⵛ, (ⵓ» et «ⵣ, >|< »).
- un rejet de caractère en affichant le caractère spécial '~' (par ex. « ⵇⵯ, ⵇ~ »

La valeur moyenne de taux d'erreurs de reconnaissance des caractères par fonte est présentée dans le tableau suivant :

Fonte d'apprentissage	**Erreurs (%)**	**Fonte de test**	**Erreur (%)**
IRCAM unicode	18	Amawas	21
tamalut	15	Anafaw	35
tamzwart	32	Azurar	15
tassafut	21	caligraph	21
Teddus	15		
taromit	9		
aggug	12		
CER	**17 %**	**CER**	**23 %**

Tableau 1 : Taux d'erreur de reconnaissance des caractères Tifinaghe par fonte

3 https://en.wikipedia.org/wiki/ Levenshtein_distance

Le tableau 1 montre les résultats de la transcription OCR pour toutes les fontes. Une différence de 6% entre le taux d'erreur de caractère pour les fontes d'apprentissage et celui des fontes de test ce qui confirme la forte tendance du moteur Tesseract vers l'indépendance par rapport aux fontes, malgré le corpus d'entrainement limité (Ait ouguengay, 2011).

Conclusion

Le logiciel de reconnaissance optique du tifinaghe présenté dans le présent travail a montré une fiabilité de 82,77% de reconnaissance et une robustesse par rapport aux polices de caractères. Néanmoins, notre futur travail sera l'augmentation du taux de fiabilité de logiciel à travers plusieurs actions envisageable à savoir : l'augmentation de la base de données d'entrainement et le post-traitement des résultats de l'OCR en enrichissant le modèle de l'alphabet tifinaghe dans le moteur de reconnaissance Tesseract.

Référence

Ait ouguengay, 2011. « Intégration de l'amazighe dans un OCR OpenSource : Ocropus comme modèle ». *Actes de la conférence SITACAM'11*, Agadir, Maroc.

Smith R. (2007). « An overview of the Tesseract OCR engine ». *Proc. Of the 9th Int. Conference on Document Analysis and Recognition (ICDAR),* IEEE Computer Society, pp. 629-633.

Rice S.V., Jenkins F.R., Nartker T.A. (1995). «The Fourth Annual Test of OCR Accuracy». Technical Report 95-03, Information Science Research Institute, University of Nevada, Las Vegas.

Mathew M., Singh A. K., Jawahar C. V.(2016). « Multilingual OCR for Indic Scripts ». *Proceedings of 12th IAPR International Workshop on Document Analysis Systems (DAS'16), 11-14 April, 2016, Santorini, Greece.*

TICAM'14, p. 219-223

La régionalisation avancée, l'officialisation de la langue amazighe quels impacts sur le développement des technologies d'information et de communication, cas de l'internet ?

Mounia Boumediane[1] **Zineb Bennouri**[2]
[1] Université Mohamed V, Faculté des lettres et sciences humaines de Rabat.
Boumediane.mounia@gmail.com
[2] Université Mohamed V, Faculté des lettres et sciences humaines de Rabat.
bennouri_zineb@hotmail.com

Résumé

Par l'adoption de la loi organique sur la régionalisation avancée et celle de l'officialisation de la langue amazighe, le législateur marocain a choisi de faire de l'arabe et de l'amazighe deux langues du pays et d'établir des droits relatifs à leur emploi.

Conscient du principe que la langue est un vecteur de communication humaine aussi bien que informatique, la langue régionale est certainement un levier du développement local dans tous les domaines. D'ici découle notre réflexion, quel est l'impact de la régionalisation avancée et l'officialisation de la langue amazighe sur le développement des technologies d'information et de communication prenant comme exemple l'internet ?

1. Introduction

Aujourd'hui au Maroc, beaucoup de foyers disposent d'un ordinateur avec une connexion Internet. Ce moyen de communication est devenu presque indispensable. Il s'agit, en effet, d'un outil formidable au niveau de la communication et de l'information. D'une part, l'Internet peut être utile pour le travail scolaire, pour effectuer des recherches, pour les devoirs, les présentations scolaires, etc. Internet n'a pas de limite et ne cesse de s'accroître.

D'autre part, l'Internet a la capacité de contribuer au changement des mentalités des utilisateurs, car en entrant en possession de ce dernier, tout internaute se transforme en acteur à part entière de la société de l'information. Il pourrait donc, bénéficier de la transparence administrative et de ses mises à jour dans divers domaines. En outre sur le plan politique, les citoyens pourraient participer plus activement à l'élaboration des politiques publiques, et le débat public ne serait que de meilleure qualité.

Cependant, ce processus est-il démocratique et participatif ? Serait-il conciliable avec les lois internationales sur l'information et avec la nouvelle Constitution marocaine ? Qui stipule que, l'information et les communications en ligne doivent être offerts selon le principe de l'égalité réelle entre tous les marocains, quelle que soient leur langue ou leur provenance régionale. Qu'elle est la prise de décision stratégique de l'Etat en la matière ?

Cette communication se portera donc, sur l'importance de la présence de la langue amazighe sur le Net, afin de constater, dans quelle mesure les nouvelles technologies d'information et de communication seraient porteuses de bénéfice. Soit au niveau de la diffusion, que par rapport à l'accès à l'information public à tout citoyen marocain, sans distinction de langue. Cela, en donnant un aperçu circonspect, sans être exhaustif, sur l'utilisation et l'exploitation universelle des NTIC, énormes outils de communication, dans l'ère de la régionalisation et l'officialisation de la langue amazighe au Maroc.

En d'autres termes, nous allons exposer l'impact sociolinguistique des NTIC et ses avantages dans la vulgarisation de la connaissance pour tous. En étant convaincues que, la combinaison entre régionalisation et technologies d'information et de communication pourrait, effectivement, mettre en exergue un ensemble d'éléments et de données utiles à la valorisation de l'alliance, du métissage identitaire et culturel. Ainsi, que la lutte contre la stigmatisation du fait de l'origine régional ou ethnique, la couleur, la langue, ou les traits et habitudes culturels.

2. Problématique

Même si, au cours des dernières années le Maroc a fait des progrès pour rétablir une large couverture Internet au niveau nationale, il reste encore des espaces, notamment les zones non arabophones, qui sont mal ou pas desservies par le réseau, où le droit à l'information n'est pas très connu du grand public car malheureusement, en dehors des personnes proches du gouvernement et des militants de la société civile, peu de personnes sont conscientes de leur droit d'accès aux informations des organismes publics ou des avantages que cela peut représenter dans leur vie quotidienne.

Lorsqu'on fait référence aux défis entourant l'utilisation d'Internet, des nouveaux médias et des médias sociaux, l'accès est au fondement de tout, pour le citoyen, le journaliste, l'élu, le parlementaire… l'accès à l'information est d'abord, l'accès à l'information publique. C'est-à-dire, à toute information qui relève du domaine public, détenue par le gouvernement et par toutes les agences d'intérêt public. Donc, il est important que le gouvernement marocain garantisse le plus vite possible un accès à tous les marocains à la large bande et aux réseaux numériques, peu importe la langue, la provenance ou le lieu de résidence.

Par ailleurs, dans le contexte sociolinguistique, les nouvelles technologies semblent prometteuses à bien des égards, essentiellement, quand il s'agit d'assurer le développement des communautés de langue officielle en situation minoritaire, poussant et favorisant la promotion de la dualité linguistique. Cela, en rendant par exemple accessibles les services public en ligne au plus grand nombre possible de personnes de différent bagage sociolinguistique. Ou encore, permettant aux fonctionnaires administratifs et aux citoyens de collaborer dans la

langue de leur choix. De ce fait, ces avancées dans le secteur des multimédias ne pourraient prendre sens que, si l'on assure un accès partout et pour tous au Maroc.

D'autant plus que, les langues ne peuvent s'empêcher de cohabiter et de se frotter les unes aux autres dans des rapports souvent complexes, créant des contacts qui peuvent être pacifiques, mais le plus souvent conflictuels. Et ce, en engendrant des phénomènes linguistiques divers, pouvant aller du simple emprunt lexical ou phonétique, à des calques syntaxiques, touchant ainsi à l'intégrité structurelle de la langue. Rappelons-nous qu'en règle générale, les langues vivent avec et à travers les communautés qui les utilisent et leur civilisations. Comme avait dit le sociolinguiste Pierre Bourdieu :« la langue ne vaut que ce que valent ses locuteurs. Si ces mêmes usagers de la langue ne produisent pas sur différents plans scientifiques, économiques, littéraires et autres…, ils sont tenus de se servir des productions des autres ».

3. Situation actuelle

Le droit d'accès à l'information a été progressivement affirmé par les organisations internationales et régionales, et il est désormais reconnu comme un droit fondamental par la jurisprudence internationale. Les fondements du droit à l'information ont été posés, en effet, par les Nations Unies dès la première session de leur Assemblée Générale, en 1946. La liberté de l'information a été alors reconnue comme « un droit de l'Homme fondamental et la pierre de touche de toutes les libertés à la défense desquelles se consacrent les Nations Unies [1]. »(Résolution 59(I)).

Ensuite, dans l'article 19, de la Déclaration universelle des droits de l'Homme de 1948, le droit à l'information est devenu une obligation contraignante avec le Pacte international relatif aux droits civils et politiques de 1966. Celui-ci dit, dans son article 19.2, que « Toute personne a droit à la liberté d'expression ; ce droit comprend la liberté de rechercher, de recevoir et de répandre des informations et des idées de toute espèce, sans considération de frontières, sous une forme orale, écrite, imprimée ou artistique, ou par tout autre moyen de son choix. ».

Enfin, c'est l'UNESCO, la seule organisation des Nations Unies, quia reçu le mandat de promouvoir la liberté de l'information. Ainsi, depuis des années, l'organisation s'est engagée à contribuer à l'amélioration des politiques publiques et des stratégies en faveur de l'accès universel à l'information. L'UNESCO s'attache surtout à promouvoir dans le monde l'accès à l'information du domaine public gouvernemental. Ses Etats membres disposent de textes de référence en la matière depuis, l'adoption par l'Organisation en 2003, de la Recommandation sur la promotion et l'usage du multilinguisme et l'accès universel au cyberespace et la publication en 2004 des Principes directeurs pour le développement et la promotion de l'information du domaine public gouvernemental.

Au Maroc, le droit à l'information public n'a été prévu qu'en juillet 2011, dans l'article 27, de la Constitution marocaine qui stipula : « Les citoyennes et les citoyens ont le droit d'accéder à l'information détenue par l'administration publique, les institutions élues et les organismes investis de mission de service public. Le droit à l'information ne peut être limité

1 http://www.un.org/documents/ga/res/1/ares1.htm.

que par la loi, dans le but d'assurer la protection de tout ce qui concerne la défense nationale, la sûreté intérieure et extérieure de l'Etat, ainsi que la vie privée des personnes, de prévenir l'atteinte aux libertés et droits fondamentaux énoncés dans la constitution ».

Droit qui advint suite aux évènements de janvier 2010, lorsque le Roi Mohammed VI avait chargé une commission consultative de penser à un modèle de régionalisation avancée proprement marocain. Ainsi, la Commission avait présenté, un peu plus d'une année plus tard, un projet de découpage du nouveau Maroc en 12 régions. Une nouvelle configuration territoriale du Maroc se profila donc, et devint effective dès l'adoption de la loi organique relative à la régionalisation avancée. Ainsi, cette nouvelle configuration, est venue rompre avec toutes les expériences de déconcentration et, plus tard, de décentralisation qu'a vécue le Maroc pendant les quatre dernières décennies.

La régionalisation implique clairement l'utilisation administrative de l'amazigh, avec l'alphabet Tifinagh. Chose qu'appuie l'article 5 de la nouvelle Constitution [2], qui prévoit, sans aucun discernement, l'intégration de cette langue dans la vie publique des régions concernées. En effet, «Il s'agit d'une initiative d'avant-garde, qui constitue le couronnement du processus de réhabilitation de l'Amazigh, comme patrimoine commun à tous les Marocains», avait affirmé le Roi Mohamed VI.

4. Questionnement

Bien qu'auparavant, parmi les moyens classiques de la diffusion et de l'épanouissement d'une langue ou au contraire à sa désintégration, l'école arrivait en force. Aujourd'hui l'Internet, servant le même dessein, est incontestablement plus efficient car son impact est encore plus important et plus intense que celui de l'enseignement. Puisqu'il atteint en même temps toutes les couches de la société et toutes les générations, des moins jeunes au plus vieux, cela dans tous les espaces, même les plus éloignés et en un temps très réduit.

L'Internet est un Outil de communication moderne et rapide, qui abolit les frontières, contourne la censure, offre la possibilité d'exercer un droit fondamentale, qu'est: la liberté d'expression. Son existence a permis d'inscrire Tamazight, identité, langue et culture, dans le processus de la mondialisation, ainsi que sa diffusion la plus large. En effet, comme nouveau moyen de technologie de l'information et de la communication, l'Internet est le moyen le plus efficient pour la vulgarisation de la culture amazighe, en témoignent la présence de quelques incontournables sites amazighs. En effet, En tapant le mot 'amazigh' ou 'berbère' dans un moteur de recherches, on se retrouve avec une foule de sites qui traitent l'amazighité sous tous ses angles.

---

2 ARTICLE 5. « L'arabe demeure la langue officielle de l'Etat. L'Etat œuvre à la protection et au développement de la langue arabe, ainsi qu'à la promotion de son utilisation. De même, l'amazighe constitue une langue officielle de l'Etat, en tant que patrimoine commun à tous les marocains sans exception. Une loi organique définit le processus de mise en œuvre du caractère officiel de cette langue, ainsi que les modalités de son intégration dans l'enseignement et aux domaines prioritaires de la vie publique, et ce afin de lui permettre de remplir à terme sa fonction de langue officielle ».

Toutefois, Internet et les sites qui vont avec, en tant que dispositif de communication moderne, qui ont la capacité de couvrir de large surface, symbolise-t-ils vraiment toutes les régions du Maroc ? Sachant que, les représentants des communautés Amazigh pensent que, la représentation de leur langue est absente ou pas à la hauteur de leurs espérances dans les sites officiels des institutions marocaines. Et, qu'elle ne correspond guère aux attentes espérées des citoyens amazighs, surtout au niveau du rapprochement et de l'amélioration des relations entre les administrations et les fonctionnaires de l'état.

Par conséquent, les NTIC, assurent-elles une bonne transparence des activités gouvernementales? ainsi qu'une plus grande participation des citoyens à la prise de décisions? Etant donné que, l'évolution vers une logique de service bilingue aux usagers, avait pour but de faire véritablement participer les tous ressortissants marocains à l'élaboration des changements et des solutions aux problèmes du pays, faisant évidemment précéder la décision par une négociation en bonne et due forme.

5. Conclusion

Les flux d'informations n'ont désormais plus de frontières, les citoyens marocains habitués à accéder à des sources d'information étrangères sur le web veulent les obtenir de leurs propres gouvernements. En effet, ces avancées technologiques ont entraînées des changements dans les comportements individuels. Le fonctionnement vertical traditionnel des administrations, fondé sur le monopole de l'information par les échelons hiérarchiques les plus élevés, a été remis en cause. Car, il ne correspond plus aux attentes de la société. Cela grâce à l'expansion des nouvelles technologies de l'information et de la communication, notamment la généralisation de l'Internet, ces dix dernières années, qui ont stimulé la demande d'information de la part du public, des entreprises et des organisations non gouvernementales.

Ainsi, dans cette tour de Babel, le Maroc entend veiller premièrement, à la cohérence de la politique linguistique et culturelle du pays, ensuite à l'apprentissage et la maîtrise des langues étrangères les plus utilisées dans le monde, en tant qu'outils de communication, d'intégration et d'interaction avec la société du savoir. Afin, de favoriser l'ouverture sur les différentes cultures et sur les civilisations contemporaines.

6. Bibliographie

Secrétariat Général du Gouvernement, *La Constitution 2011*, série « documentation juridique marocaine », imprimerie officielle, édition 2011.

Pierre Bourdieu, le marché des biens symboliques, l'année sociologique, 3e série, vol. 22, 1971

7. Web graphie

Rapport sur la régionalisation avancée, http://www.regionalisationavancee.ma/PDF/Rapport.

RESOLUTIONS ADOPTED BY THE GENERAL ASSEMBLY DURING ITS FIRST SESSION http://www.un.org/documents/ga/res/1/ares1.htm.

Vers les societies du savoir, http://unesdoc.unesco.org.